Open Archive TOULOUSE Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: http://oatao.univ-toulouse.fr/
Eprints ID: 1655

To cite this version: Vasseur, Xavier *Contribution to the study of efficient iterative methods for the numerical solution of partial differential equations.* (2016) [HDR]

Any correspondence concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr

Habilitation à Diriger des Recherches

Spécialité : Mathématiques appliquées

Contribution to the study of efficient iterative methods for the numerical solution of partial differential equations.

Contribution à l'étude de méthodes itératives efficaces pour la résolution numérique des équations aux dérivées partielles.

proposée au sein de l'Ecole Doctorale de Mathématiques, Informatique et Télécommunications de Toulouse
(Institut National Polytechnique de Toulouse)

par

Xavier Vasseur

CERFACS

soutenue le 1er juin 2016 devant le jury composé de :

Prof. Iain S. Duff	Rutherford Appleton Laboratory, UK and CERFACS, France	Examiner
Prof. Andreas Frommer	University of Wuppertal, Germany	Referee
Prof. Serge Gratton	INPT-IRIT, France	Examiner
Frédéric Nataf	Université Pierre et Marie Curie, France	Referee
Prof. Cornelis Oosterlee	CWI and TU Delft, The Netherlands	Referee
Michel Visonneau	Ecole Centrale de Nantes, France	Examiner

après avis des rapporteurs :

Prof. Andreas Frommer	University of Wuppertal, Germany	Referee
Frédéric Nataf	Université Pierre et Marie Curie, France	Referee
Prof. Cornelis Oosterlee	CWI and TU Delft, The Netherlands	Referee

Abstract

Abstract Multigrid and domain decomposition methods provide efficient algorithms for the numerical solution of partial differential equations arising in the modelling of many applications in Computational Science and Engineering. This manuscript covers certain aspects of modern iterative solution methods for the solution of large-scale problems issued from the discretization of partial differential equations. More specifically, we focus on geometric multigrid methods, non-overlapping substructuring methods and flexible Krylov subspace methods with a particular emphasis on their combination. Firstly, the combination of multigrid and Krylov subspace methods is investigated on a linear partial differential equation modelling wave propagation in heterogeneous media. Secondly, we focus on non-overlapping domain decomposition methods for a specific finite element discretization known as the *hp* finite element, where unrefinement/refinement is allowed both by decreasing/increasing the step size h or by decreasing/increasing the polynomial degree p of the approximation on each element. Results on condition number bounds for the domain decomposition preconditioned operators are given and illustrated by numerical results on academic problems in two and three dimensions. Thirdly, we review recent advances related to a class of Krylov subspace methods allowing variable preconditioning. We examine in detail flexible Krylov subspace methods including augmentation and/or spectral deflation, where deflation aims at capturing approximate invariant subspace information. We also present flexible Krylov subspace methods for the solution of linear systems with multiple right-hand sides given simultaneously. The efficiency of the numerical methods is demonstrated on challenging applications in seismics requiring the solution of huge linear systems of equations with multiple right-hand sides on parallel distributed memory computers. Finally, we expose current and future prospectives towards the design of efficient algorithms on extreme scale machines for the solution of problems coming from the discretization of partial differential equations.

Keywords Algebraic multigrid method (AMG); Augmentation; Balancing Neumann-Neumann (BNN); Block Krylov subspace method; Block size reduction; Deflation; Finite Element Tearing and Interconnecting (FETI); Flexible Krylov subspace method; Full Approximation Scheme (FAS); Full Multigrid (FMG); Helmholtz equation; High Performance Computing (HPC); *hp* finite element method; Iterative method; Krylov subspace method; Linear systems of equations with multiple right-hand sides; Multigrid method; Non-overlapping domain decomposition method; Preconditioning; Spectral deflation; Substructuring method; Variable preconditioning.

Acknowledgments

This manuscript summarizes my main research activities in numerical analysis and scientific computing. This research benefited from a large number of people whom I wish to thank.

Firstly I am very much indebted to Prof. Serge Gratton, who helped me developing my research activities at CERFACS. I must sincerely thank him for his friendship, generosity, continuous and very valuable support in every circumstance, his numerous advices and ideas that have made this long term collaboration successful. I am looking forward to pursuing this fruitful collaboration in the near future. Thanks for all, Serge !

Secondly I would like to thank Prof. Andreas Frommer, Prof. Cornelis Oosterlee and Dr. Frédéric Nataf for acting as referees and for participating to the jury in Toulouse. Their in depth reading of the manuscript, reports and numerous questions during the defence were highly appreciated. Dr. Michel Visonneau is also particularly acknowledged for his continuous interest in iterative methods and for taking in charge the presidence of the jury. It has been a real honour for me. Finally, I would like to thank Prof. Iain S. Duff for his careful reading of the manuscript and his interest in the research activities that I have been able to develop at CERFACS. I have really appreciated the insight, expertise and kindness of all the jury members.

Brigitte Yzel deserves a special thanks for the administrative support and the organization of the defence on June 1st.

Before arriving at CERFACS, I have had the opportunity to contribute to research in different outstanding groups in the field of fluid mechanics, numerical analysis and scientific computing, each group being stimulating in their own way. Thus I am very grateful to Prof. Jean Piquet, Dr. Michel Visonneau, Prof. Alfio Quarteroni, Dr. Andrea Toselli and Prof. Christoph Schwab for their support, confidence and guidance over the past years.

As already emphasised, this research benefited from many collaborations. Hence I would like to thank all the students, coauthors, colleagues and industrial partners I have worked with during these past years. Special thanks go to past and current colleagues in the Parallel Algorithms Project.

Finally, I would like to thank my parents and relatives for their continuous and invaluable support. Last, but by no means least, I am greatly indebted to Caroline, Constance and Quentin for their love, support and understanding.

Toulouse, June 2016.

Contents

List of Figures

List of Tables

List of Algorithms

Notation

$\mathbb{N}$, $\mathbb{R}$, $\mathbb{C}$	Set of non-negative integers, real numbers and complex numbers.
$\mathcal{V} \oplus \mathcal{W}$	Direct sum of two subspaces $\mathcal{V}, \mathcal{W} \subseteq \mathcal{H}$ with $\mathcal{V} \cap \mathcal{W} = \{0\}$.
I_n	Identity matrix $I_n \in \mathbb{C}^{n \times n}$.
L^{H}, L^T	Hermitian transpose matrix and transpose matrix.
$\mathsf{P}_{\mathcal{V},\mathcal{W}}$	Projection onto $\mathcal{V}$ along $\mathcal{W}$.
$\mathcal{N}(\mathsf{L})$	Null space of L.
$\mathcal{R}(\mathsf{L})$	Range of L.
$\Lambda(\mathsf{L})$	Spectrum of L.
$\Sigma(\mathsf{L})$	Singular values of L.
$\kappa(\mathsf{L})$	Condition number of L: $\kappa(\mathsf{L}) = \|\mathsf{L}\| \left\|\mathsf{L}^{-1}\right\|$.
$\|.\|_2$	Euclidean norm.
$\|.\|_F$	Frobenius norm.

1. Introduction

Scope and goals

Scope

Computational Science and Engineering (CSE) is a multidisciplinary field aiming at simulating complex phenomena by exploiting the power of modern computational resources. In this respect, the simulation of physical phenomena governed by nonlinear and time-dependent partial differential equations (PDEs) plays a major role. Beyond simulation, current CSE research projects focus more and more on the optimization or on the control of those complex physical phenomena governed by PDEs. Hence, efficient solution methods for the numerical solution of partial differential equations must be provided. In this manuscript, we consider standard discretization methods for the numerical approximation of deterministic PDEs ranging from finite difference, finite volume or finite element methods (in both h and hp versions, respectively). Since the problems considered are often of multiscale nature, it is highly relevant to represent the different scales of the physical phenomena. Consequently, the discretization of the continuous equations usually reduces the problem to the solution of an huge nonlinear algebraic system of equations. In this setting, we will exclusively consider efficient and numerically stable parallel algorithms based on *iterative* methods for the solution of such algebraic systems. This is the main scope of the manuscript, where analysis and performance of such algorithms will be specifically examined on applications related to fluid dynamics, geophysics and structural mechanics.

This scope involves multiple areas of research. Theoretical aspects in functional analysis and calculus of variations are needed to analyse questions related to the existence, unicity and behavior of solutions of PDEs. Furthermore, the discretization of infinite dimensional problems requires knowledge of function spaces and approximation theory. Finally, the solution of discretized finite dimensional problems is investigated through matrix computations, where efficient and numerically stable algorithms are favoured on massively parallel architectures. In this manuscript, we mostly concentrate on the algebraic iterative computational aspects but we are certainly aware that all these topics are interconnected; see [153, 171, 176, 199] for enlightening discussions on this aspect.

Optimality and scalability of iterative methods

We refer the reader to standard textbooks on iterative methods to discover the plethora of available algebraic methods [15, 24, 133, 140, 180, 194, 217, 257]. The central notion of preconditioning is addressed in the monographs [90, 194, 217, 253, 257] and in the

survey papers [27, 262]. We next define *optimality* and *scalability* of iterative methods, two terms that are frequently used throughout the manuscript; see [246, Chapter 1, Definitions 1.2 and 1.3].

Definition 1.1. Optimality. An iterative method for the solution of a linear system is said to be optimal, if its rate of convergence to the exact solution is independent of the size of the system.

Here, optimality is ensured if the rate of convergence is independent of the size of the finite element space employed (meshsize h for h approximations and polynomial degree for hp approximations) or of the meshsize h for finite difference or finite volume discretizations.

To introduce parallelism when considering mesh-based simulations of PDEs, the global computational domain in space is usually partitioned into smaller subdomains [226]. A one-to-one mapping between processors and subdomains is then used in the standard implementation of the numerical methods.

Definition 1.2. Scalability. An iterative method for the solution of a linear system is said to be scalable, if its rate of convergence does not deteriorate when the number of subdomains grows.

Definitions 1.1 and 1.2 are given in the context of the solution of linear systems of equations and can be easily extended to the nonlinear case.

Goals

The overall objective in the manuscript is the analysis and development of advanced numerical methods for the simulation of PDE-based applications that are able to efficiently exploit the power of parallel computers. In this regard, we have in mind two specific goals

- **Analysis and implementation of optimal or scalable linear solvers for systems arising from mesh-based implicit simulation of PDEs.**

  This is the major key step when solving PDEs implicitly. Depending on the nature of the PDEs, *optimal* or *scalable* solvers may be designed. Finally, we note that the situation is much more intricate, when systems of PDEs are considered.

- **Analysis and implementation of optimal or scalable algorithms beyond linear solvers.**

  The partial differential equations of interest are often nonlinear and possibly time-dependent. Hence, the design of *optimal* or *scalable* algorithms for nonlinear systems of equations is of utmost importance in practice. Furthermore, the time

variable may offer an additional possibility for introducing parallelism. If successful, this would lead to optimal or scalable algorithms that are parallel in both time and space.

To reach these goals, we consider the combination of multilevel methods (of either geometric multigrid or domain decomposition type) and of Krylov subspace methods to design efficient numerical algorithms for the solution of large-scale problems coming from the discretization of partial differential equations. We refer the reader to the monographs [15, 133, 166, 194, 217, 257] and the comprehensive survey paper [224] for details on theoretical and practical aspects of Krylov subspace methods. In the manuscript, we will pay a specific attention to the numerical properties of the combination of Krylov subspace methods and multilevel preconditioners. We briefly study the most salient properties of those multilevel preconditioners next.

Geometric multigrid methods are known to be *optimal* iterative methods for certain classes of discretized elliptic PDEs [39, 40, 141]. In the case of a Laplace operator with constant coefficients, the full multigrid method (FMG) [39, 234] is considered to be asymptotically optimal, that is, the number of arithmetic operations required is proportional to the number of grid points, with only a small constant of proportionality; see [40, Chapter 7], [247, Appendix C] and [208] for the corresponding analysis. In addition, the Full Approximation Scheme (or Full Approximation Storage) (FAS) [39] and [40, Chapter 8] has been demonstrated to be an effective nonlinear multilevel method [141] for the solution of discretized partial differential equations. Hence, multigrid methods provide particularly relevant algorithms to consider in our framework. For further details, we refer the reader to the standard monographs [40, 45, 141, 177, 234, 247] for the mathematical analysis of linear and nonlinear geometric multigrid methods. Parallelization of geometric multigrid methods is especially discussed in [247, Chapter 6].

Domain decomposition refers to the splitting of a partial differential equation into coupled problems on smaller subdomains forming a partition of the original computational domain [204, 226]. This splitting can be performed at the continuous, discrete level or at the algebraic level. While parallelism is natural due to the domain partitioning, the key question in domain decomposition is how to select the subproblems to ensure that the rate of convergence of the iterative method is fast. We consider iterative substructuring domain decomposition methods equipped with a coarse space, that are known to provide *scalable* algorithms for the solution of linear elliptic partial differential equations. We refer the reader to the monographs [80, 204, 226, 246] for historical comments and detailed analysis of various domain decomposition methods. The abstract theory of Schwarz methods (the earliest domain decomposition algorithm is the alternating Schwarz method) is presented in [246, Chapters 2 and 3], while an algebraic theory has been proposed in [29, 111, 113]. Finally, we mention the review article by Xu [267] and the book chapter by Oswald [247, Appendix B] for a description of an abstract theory of multilevel methods in terms of subspace decomposition.

Outline

The remainder of the manuscript is divided into four chapters.

Chapter 2 is related to the study of geometric multigrid methods for the solution of the Helmholtz equation in three-dimensional heterogeneous media. This is known as a difficult problem for iterative methods [100] and *optimal* solvers have not been proposed yet. We analyse the combination of geometric multigrid preconditioners and Krylov subspace methods in this setting. Then we illustrate the main properties of the resulting numerical methods on a realistic application in exploration seismology requiring the solution of linear systems of billion of unknowns in practice. Scalability properties are investigated on massively parallel computers.

Chapter 3 is related to the study of domain decomposition preconditioners for *hp* finite element approximations on anisotropic meshes in two and three dimensions. When simulating physical phenomena exhibiting boundary layers or singularities, geometrically refined meshes towards corners, edges or faces must be used in a *hp* finite element formulation. In consequence, two- or three-dimensional meshes with high aspect ratios have to be employed in practice. Hence, the condition number of the stiffness matrix rapidly deteriorates: it grows exponentially with the spectral polynomial degree k. The solution of such linear systems with iterative methods is thus especially difficult. We focus on two non-overlapping domain decomposition preconditioners known as Balancing Neumann-Neumann and FETI, respectively. We give condition number bounds for the preconditioned operators and prove that the proposed numerical methods are *scalable* in such a context. Numerical experiments supporting this conclusion are studied in detail on academic partial differential equations in two and three dimensions.

Chapters 2 and 3 have provided several examples of variable multilevel preconditioners (i.e. the preconditioner is not a fixed linear operator). These preconditioners must be used with a specific class of Krylov subspace methods named flexible Krylov subspace methods. In addition to preconditioning, it is known that deflation and augmentation are two features that can improve the rate of convergence of Krylov subspace methods. Hence, in Chapter 4, we propose and analyse flexible Krylov subspace methods combining spectral deflation and/or augmentation. We also derive advanced flexible Krylov subspace methods for the solution of linear systems with multiple right-hand sides given simultaneously. The efficiency of the numerical methods is finally demonstrated on challenging large-scale applications in seismics requiring the solution of huge linear systems of equations with multiple right-hand sides on parallel distributed memory computers.

In Chapter 5, we briefly explore prospectives towards the numerical solution of deterministic or stochastic partial differential equations on future computing platforms. Indeed, new algorithms should be designed to be able to exploit as efficiently as possible the power of extreme scale computers. We address a few research prospectives on

both multilevel preconditioners and Krylov subspace methods. Current prospectives are illustrated on applications related to porous media flows in reservoir modelling or structural mechanics.

The manuscript ends with two appendices and references.

How to read the manuscript

Chapters 2, 3 and 4 aim at providing an overview of the main results that have been obtained so far. To go further in the analysis, the reader is referred first to the five selected publications proposed in Appendix B, and then to the bibliography, respectively. Emphasis is made on multilevel preconditioners and Krylov subspace methods, while material related to partial differential equations can be found in reference textbooks.

2. Geometric multigrid methods for three-dimensional heterogeneous Helmholtz problems

2.1. Objectives and contributions

2.1.1. Objectives

In this chapter, we focus on geometric multigrid methods for the solution of linear systems of equations arising from the discretization of partial differential equations. We exclusively concentrate on a specific partial differential equation, the Helmholtz equation written in the frequency domain, modelling acoustic wave propagation phenomena in an infinite medium. We are especially interested in solving wave propagation phenomena in three-dimensional heterogeneous media, as required in applications such as exploration seismology (oil exploration, earthquake modelling) and acoustic scattering.

Large scale heterogeneous Helmholtz problems are notoriously difficult to solve; see [96, 100] for comprehensive surveys. Hence, providing robust iterative solution methods with respect to both the mesh size and the frequency is still an open question, despite the numerous attempts in the applied mathematics community related to, e.g., sparse direct methods, domain decomposition or multigrid methods. In this chapter, our objectives are twofold

- to propose and analyse geometric multigrid methods used as preconditioners of Krylov subspace methods for the solution of large-scale linear systems arising in this setting,
- to provide detailed numerical experiments focusing on the scalability properties of the resulting numerical methods on massively parallel computers.

2.1.2. Contributions

The main contributions presented in this chapter are

- a multilevel extension of the geometric two-grid preconditioner for the solution of three-dimensional heterogeneous Helmholtz problems proposed in [58] (also given in Appendix B.4),
- a brief analysis of the resulting multilevel preconditioner based on rigorous Fourier analysis given in [58],

- a detailed report on numerical experiments, when discretizations based on either second order or high order finite difference schemes are used.

These contributions concern algorithmic, theoretical and computational aspects of multigrid methods, respectively.

Past contributions For the sake of brevity, we address a single (and currently challenging) topic in this chapter. Geometric multigrid methods for elliptic scalar PDEs or nonlinear systems of PDEs have been considered by the author in the past. Indeed, in his PhD thesis [252], the author proposed geometric multigrid methods in two and three dimensions for the solution of linear systems with symmetric positive semidefinite matrices in the context of Computational Fluid Dynamics [201, 202]. In addition, he also proposed a nonlinear geometric multigrid method based on the Full Approximation Scheme (FAS) [39] for the fully coupled solution of the incompressible Navier-Stokes equations in three dimensions [252, Section 3.10] (see also [251] and [76] (given in Appendix B.1) for the derivation of the discrete formulation). We refer the reader to [201, 202, 252] for a detailed description of these numerical methods, where numerical experiments on both academic and realistic problems in fluid mechanics are reported showing the efficiency of the linear and nonlinear multigrid solvers. *Optimal* solvers in the sense of Definition 1.1 (i.e. the number of iterations is found to be independent of the number of unknowns) were designed for such applications in Computational Fluid Dynamics on both sequential and vector computing platforms at that time.

2.1.3. Specific notation

Classical generic notation specific to the multigrid setting is introduced here. Given a physical domain Ω, the fine and coarse discrete levels denoted by h and H are associated with discrete grids Ω_h and Ω_H, respectively. In geometric multigrid methods, a geometric construction of the coarse grid Ω_H is considered. The discrete coarse grid domain Ω_H is then obtained from the discrete fine grid domain Ω_h by doubling the mesh size in each direction, as is standard in vertex-centered geometric multigrid [234]. Given a continuous operator A defined on Ω, we assume that A_H represents a suitable approximation of the fine grid operator A_h on Ω_H. We also introduce $I_h^H : \mathcal{G}(\Omega_h) \to \mathcal{G}(\Omega_H)$ a restriction operator, where $\mathcal{G}(\Omega_k)$ denotes the set of grid functions defined on Ω_k. Similarly $I_H^h : \mathcal{G}(\Omega_H) \to \mathcal{G}(\Omega_h)$ will represent a given prolongation operator. More precisely, we select as a prolongation operator trilinear interpolation and as a restriction its adjoint which is often called the full weighting operator [234]. We refer the reader to [247, Section 2.9] for a complete description of these operators in three dimensions.

2.1.4. Synopsis

We first propose a selective (and therefore incomplete) literature review related to numerical methods for the simulation of wave propagation in heterogeneous media in Section 2.2. In Section 2.3, we briefly describe the continuous and discrete settings.

Then, we exclusively focus on multigrid methods used as preconditioners for the solution of large scale Helmholtz problems. We first describe in detail the complex shifted Laplace preconditioner (Section 2.4) and then introduce the two-grid preconditioners that have been proposed (Sections 2.5 and 2.6, respectively). In Section 2.7 we analyse by rigorous Fourier analysis the main properties of the improved two-grid preconditioner and illustrate its salient properties on a realistic three-dimensional problem in geophysics in Section 2.8. Conclusions are proposed in Section 2.9.

2.2. Literature review

In a finite element setting, it is known that the standard variational formulation of the Helmholtz equation is sign-indefinite (i.e. not coercive) [181]. This means that it is difficult to find error estimates for the Galerkin method that are explicit in the wavenumber, and to prove anything a priori about how iterative methods behave when solving the Galerkin linear system. The literature on iterative solvers for discrete Helmholtz problems is thus quite rich and we refer the reader to the survey papers [96, 100] for a taxonomy of advanced preconditioned iterative methods based on domain decomposition or multigrid methods. We briefly comment on a few references amongst this rich literature.

Advanced sparse direct solvers Recent advances in sparse direct methods based on Gaussian elimination (multifrontal methods) have allowed the efficient treatment of large matrices; see, e.g., the monographs [73, 84]. Factorization based on low-rank approximation or hierarchically structured solvers have been designed to considerably lower the complexity in both factorization and solution phases; see, e.g., [259, 260]. In such a setting, applications to the Helmholtz equation in heterogeneous media have been provided in [10, 263] in the context of the MUMPS software [7, 8, 9]. In this family, we mention the moving PML sweeping method [203] based on a block incomplete LDU factorization. Hierarchical $\mathcal{H}$-matrix compression techniques are key aspects in this algorithm leading to attractive complexities. If N denotes the total number of unknowns, the computation of the preconditioner requires $\mathcal{O}(N^{\alpha})$ operations with $\alpha > 1$, whereas the action of the preconditioner requires $\mathcal{O}(N log\ N)$ operations. This is the first method with this property.

Domain decomposition methods We refer the reader to [246, Section 11.5.2] for an excellent review of domain decomposition preconditioners for the solution of Helmholtz problems. We note that, in the case of homogeneous media, the FETI-H non-overlapping domain decomposition method [101], a generalization of the FETI method [103] discussed in Section 3.4 for Helmholtz type problems, exhibits a rate of convergence that is independent of the fine grid step size, the number of subdomains, and the wavenumber. The case of heterogeneous media is of course much more complex, as expected. Recent advances related to non-overlapping domain decomposition methods have been proposed; the research has focused on the design of optimized interface (or transmis-

sion) conditions; see, e.g., [117, 232, 254]. In [232], Stolk proposed a rapidly converging domain decomposition method with transmission conditions based on the perfectly matched layer that leads to a numerical method with a near linear complexity. The method is *scalable* in the sense of Definition 1.2, i.e., the number of iterations is essentially independent of the number of subdomains.

Multigrid methods When the medium is homogeneous (or similarly when the wavenumber is uniform), efficient multigrid solvers have been proposed in the literature. We mention the wave-ray multigrid method [41] which exploits the structure of the error components that standard multigrid methods fail to eliminate [43]. In this chapter, we prefer to focus on the case of three-dimensional Helmholtz problems defined in heterogeneous media for which the design of iterative methods that are robust with respect to the frequency for such indefinite problems is currently an active research topic.

In [25] Bayliss et al. considered preconditioning the Helmholtz operator with a different operator. A few iterations of the symmetric successive over-relaxation method were then used to approximately invert a Laplace preconditioner. Later this work was generalized by Magolu Monga Made et al [170] and Laird and Giles [163], who proposed a Helmholtz preconditioner with a positive sign in front of the Helmholtz term. In [95, 99] Erlangga et al. further extended this idea: a modified Helmholtz operator with a complex wavenumber (i.e., where a complex term (hereafter named complex shift) multiplies the square of the wavenumber) was used as a preconditioner of the Helmholtz operator. This preconditioning operator is referred to as a complex shifted Laplace operator in the literature. This idea has received a lot of attention over the last few years; see among others [98, 99, 123]. Indeed, with an appropriate choice of the imaginary part of the shift, standard multigrid methods can be applied successfully, i.e., the convergence of the multigrid method as a solver or as a preconditioner applied to a complex shifted Laplace operator is mathematically found to be mesh independent at a given frequency [207]. Nevertheless, when a multigrid method applied to a shifted Laplace operator is considered as a preconditioner for the Helmholtz operator, the convergence is found to be frequency dependent as observed in [37, 207]. This behaviour occurs independently of the way the preconditioner is inverted (approximately or exactly). A linear increase in preconditioner applications versus the frequency is usually observed on three-dimensional problems in heterogeneous media. In practice, preconditioning based on a complex shifted Laplace operator is considered nowadays as a successful algorithm for low to medium range frequencies.

At high frequency (or equivalently at large wavenumbers), numerical results on the contrary show a steep increase in the number of outer iterations (see, e.g., [207] for a concrete application in seismic imaging). The analysis of the shifted Laplace preconditioned operator provided in [123] has indeed shown that the smallest eigenvalues in modulus of the preconditioned operator tend to zero as the wavenumber increases. Hence, it becomes essential to combine this preconditioner with deflation techniques to yield an efficient numerical method as analysed in [97, 222]. As far as we know, the resulting algorithms have not yet been applied to concrete large-scale applications on realistic three-dimensional heterogeneous problems. This is indeed a topic of current

research most likely due to the complexity of the numerical method. Alternatives are required and will be proposed in this chapter.

Two recent approaches First we would like to mention the recent approach proposed by Zepeda-Núñez and Demanet [270] based on the combination of domain partitioning and integral equations with application to two-dimensional acoustic problems. The method decomposes the domain into layers, and uses transmission conditions in boundary integral form to explicitly define polarized traces, i.e., up- and down-going waves sampled at interfaces. The method exhibits an online runtime of $\mathcal{O}(N/P)$ in two dimensions, where N is the number of degrees of freedom and P is the number of nodes, in a distributed memory environment, provided that $P = \mathcal{O}(N^{1/8})$. A low number of Krylov subspace iterations is obtained on realistic two-dimensional heterogeneous problems *independently* of the frequency, making this algorithm very competitive. As far as we know, this method has not yet been extended to three-dimensional problems. Secondly, Liu and Ying [168, 169] have proposed enhancements of the sweeping preconditioner leading to a $\mathcal{O}(N)$ complexity for both the setup phase and the preconditioner application with numerical experiments in three dimensions.

2.3. Problem setting

We specify the continuous and discrete versions of the heterogeneous Helmholtz problem that we consider throughout this chapter.

2.3.1. Mathematical formulation at continuous level

Given a three-dimensional physical domain Ω_p of parallelepiped shape, the propagation of a wavefield in a heterogeneous medium can be modelled by the Helmholtz equation written in the frequency domain [238]

$$-\sum_{i=1}^{3}\frac{\partial^2 u}{\partial x_i^2}-\frac{(2\pi f)^2}{c^2}u \quad = \quad \delta(\mathbf{x}-\mathbf{s}), \quad \mathbf{x}=(x_1,x_2,x_3)\in\Omega_p. \tag{2.1}$$

In equation (2.1), the unknown u represents the pressure wavefield in the frequency domain, c the acoustic-wave velocity in ms^{-1}, which varies with position, and f the frequency in Hertz. The source term $\delta(\mathbf{x}-\mathbf{s})$ represents a harmonic point source located at $\mathbf{s}=(s_1,s_2,s_3)\in\Omega_p$. The wavelength λ is defined as $\lambda = c/f$ and the wavenumber as $2\pi f/c$. A popular approach - the Perfectly Matched Layer formulation (PML) [30, 31] - has been used in order to obtain a satisfactory near boundary solution, without many artificial reflections. Artificial boundary layers are then added around the physical domain to absorb outgoing waves at any incidence angle as shown in [30]. We denote by Ω_{PML} the surrounding domain created by these artificial layers. This formulation leads to the following set of coupled partial differential equations with homogeneous

Dirichlet boundary conditions imposed on Γ, the boundary of the domain

$$-\sum_{i=1}^{3}\frac{\partial^2 u}{\partial x_i^2}-\frac{(2\pi f)^2}{c^2}u = \delta(\mathbf{x}-\mathbf{s}) \quad \text{in} \quad \Omega_p, \tag{2.2}$$

$$-\sum_{i=1}^{3}\frac{1}{\xi_{x_i}(x_i)}\frac{\partial}{\partial x_i}\left(\frac{1}{\xi_{x_i}(x_i)}\frac{\partial u}{\partial x_i}\right)-\frac{(2\pi f)^2}{c^2}u = 0 \quad \text{in} \quad \Omega_{PML}\backslash\Gamma, \tag{2.3}$$

$$u = 0 \quad \text{on} \quad \Gamma, \tag{2.4}$$

where the one-dimensional ξ_{x_i} function represents the complex-valued damping function of the PML formulation in the i-th direction, selected as in [195]. The set of equations (2.2, 2.3, 2.4) defines the forward problem related to acoustic imaging in geophysics that will be considered in this chapter. We note that the proposed numerical method can be applied to other application fields, where wave propagation phenomena appear as well.

2.3.2. Mathematical formulation at a discrete level

Second-order finite difference scheme

As frequently used in the geophysics community, we have considered a standard second-order accurate seven-point finite difference discretization of the Helmholtz problem (2.2, 2.3, 2.4) on a uniform equidistant Cartesian grid of size $n_x \times n_y \times n_z$ (see [200, Appendix A] for a complete description of the discretization). We denote later by h the corresponding mesh grid size, Ω_h the discrete computational domain and n_{PML} the number of points in each PML layer. A fixed value of $n_{PML} = 10$ is used hereafter. Since a stability condition has to be satisfied to correctly represent the wave propagation phenomena [66], we consider a standard second-order accurate discretization scheme with 10 points per wavelength. This implies that the mesh grid size h and the minimum wavelength in the computational domain must satisfy the following inequality [66]

$$\frac{h}{\min_{(x_1,x_2,x_3)\in\Omega_h}\lambda(x_1,x_2,x_3)} \leq \frac{1}{10}.$$

Hereafter, we have considered the following relation to determine the step size h, given a certain frequency f and an heterogeneous velocity field c

$$h = \frac{\min_{(x_1,x_2,x_3)\in\Omega_h} c(x_1,x_2,x_3)}{10\,f}. \tag{2.5}$$

Dispersion minimizing finite difference scheme

Since standard second-order finite difference schemes are often found to be too dispersive [66], we have considered dispersion minimizing finite difference schemes. These schemes are especially recommended when targeting the solution of heterogeneous Helmholtz problems at high frequency, since they provide a pollution-free solution [65, 195, 233, 248]. In the context of multilevel algorithms, these schemes are also relevant for the

discretization of the coarse grid operator in order to provide the same dispersion level on both the coarse and fine scales [233]. This feature has also been found beneficial by several authors, see, e.g., [65, 233, 249]. Hereafter, we have considered the compact finite difference scheme proposed by Harari and Turkel [143] based on Padé approximations, which leads to a finite difference discretization with a 27 point stencil in three dimensions. This scheme is formally third-order accurate on general Cartesian grids and fourth-order accurate on uniform grids. Following [25], given reference values for both the frequency f_{ref} and the step size h_{ref} and denoting by q the discretization order of the finite difference scheme, we have used the following condition to determine the step size h, given a certain frequency f

$$h^q \; f^{q+1} = h_{ref}^q \; f_{ref}^{q+1}. \tag{2.6}$$

Properties of the discrete linear system

The discretization of the forward problem (2.2, 2.3, 2.4) with finite difference schemes leads to the following linear system $A_h\, x_h = b_h$, where $A_h \in \mathbb{C}^{n_h \times n_h}$ is a sparse complex matrix which is non Hermitian and non symmetric due to the PML formulation [31, 200, 228] and where $x_h, b_h \in \mathbb{C}^{n_h}$ represent the discrete frequency-domain pressure field and source, respectively. In addition, the right-hand side is usually very sparse. The conditions (2.5 or 2.6) require solving large systems of equations at the (usually high) frequencies of interest for the geophysicists, a task that may be too memory expensive for standard [228, 229] or advanced sparse direct methods exploiting hierarchically semi-separable structure [259, 260] on a reasonable number of cores of a parallel computer (see Section 2.2). Consequently, preconditioned Krylov subspace methods are most often considered and efficient preconditioners must be developed for such problems. We describe next in detail three preconditioners that have been proposed for the solution of the forward problem related to acoustic imaging.

2.4. Complex shifted Laplace multigrid preconditioner

We briefly present a popular preconditioner for the Helmholtz equation, since it will serve as a basis for the method presented in Section 2.6.

2.4.1. Algorithm and components

In [98, 99] Erlangga et al. have exploited the pioneering idea to define a preconditioning operator based on a different partial differential equation for which a truly multilevel solution is possible. The corresponding set of equations reads as

$$-\sum_{i=1}^{3} \frac{\partial^2 u}{\partial x_i^2} - (1+i\beta)\frac{(2\pi f)^2}{c^2} u \;=\; \delta(\mathbf{x}-\mathbf{s}) \quad \text{in} \quad \Omega_p, \tag{2.7}$$

$$-\sum_{i=1}^{3} \frac{1}{\xi_{x_i}(x_i)} \frac{\partial}{\partial x_i}\Big(\frac{1}{\xi_{x_i}(x_i)} \frac{\partial u}{\partial x_i}\Big) - (1+i\beta)\frac{(2\pi f)^2}{c^2} u \;=\; 0 \quad \text{in} \quad \Omega_{PML}\backslash\Gamma, \tag{2.8}$$

$$u = 0 \quad \text{on} \quad \Gamma, \tag{2.9}$$

where the parameter $1 + i\beta \in \mathbb{C}$ is called the complex shift[1]. We introduce a sequence of l grids denoted by $\Omega_1, \cdots, \Omega_l$ (with Ω_l as the finest grid) and of appropriate operators $S_k^{(\beta)}$ $(k = 1, \cdots, l)$. Here $S_k^{(\beta)}$ is simply obtained from the finite difference discretization of (2.7, 2.8, 2.9) on Ω_k. $S_k^{(\beta)}$ is later called the complex shifted Laplace operator on Ω_k. In order to describe the algorithm in detail, we denote by $I_k^{k-1} : \mathcal{G}(\Omega_k) \to \mathcal{G}(\Omega_{k-1})$ a restriction operator from Ω_k to Ω_{k-1}, $I_{k-1}^k : \mathcal{G}(\Omega_{k-1}) \to \mathcal{G}(\Omega_k)$ a prolongation operator from Ω_{k-1} to Ω_k and C the cycling strategy (which can be of V, F or W type). The complex shifted multigrid algorithm is then sketched in Algorithm 2.1. An illustration is depicted in Figure 2.1.

Algorithm 2.1 Multigrid cycle (with a hierarchy of l grids) applied to $S_l^{(\beta)} y_l = w_l$. $y_l = \mathcal{M}_{l,C}(w_l)$.

Input: Assume that the following is given

- $S_k^{(\beta)} \in \mathbb{C}^{n_k \times n_k}$ $(k = 1, \cdots, l)$ ▷ complex shifted Laplace operators discretized on Ω_k
- $w_l \in \mathbb{C}^{n_l}$ ▷ right-hand side given on Ω_l
- $y_l \in \mathbb{C}^{n_l}$ ▷ initial guess given on Ω_l

1: Pre-smoothing: Apply ν_β iterations of ω_l-Jacobi to $S_l^{(\beta)} y_l = w_l$ to obtain the approximation $y_l^{\nu_\beta}$.

2: Restrict the fine level residual: $w_{l-1} = I_l^{l-1}(w_l - S_l^{(\beta)} y_l^{\nu_\beta})$.

3: Solve approximately the coarse problem $S_{l-1}^{(\beta)} y_{l-1} = w_{l-1}$ with initial approximation $y_{l-1}^0 = 0_{l-1}$: Apply recursively γ cycles of multigrid to $S_{l-1}^{(\beta)} y_{l-1} = w_{l-1}$ to obtain the approximation y_{l-1}. On the coarsest level ($l = 1$) apply ϑ_β cycles of GMRES(m_β) preconditioned by ν_β iterations of ω_1-Jacobi to $S_1^{(\beta)} y_1 = w_1$ as an approximate solver.

4: Perform the coarse level correction: $\widetilde{y_l} = y_l^{\nu_\beta} + I_{l-1}^l y_{l-1}$.

5: Post-smoothing: Apply ν_β iterations of ω_l-Jacobi to $S_l^{(\beta)} y_l = w_l$ with initial approximation $\widetilde{y_l}$ to obtain the final approximation y_l.

In Algorithm 2.1, the γ parameter controls the type of cycling strategy of the multigrid hierarchy, see, e.g., [234]. Trilinear interpolation and full-weighting are used as prolongation and restriction operators, respectively. An approximate solution on the coarsest level is considered as in the two-grid approach proposed next in Section 2.5. We note that the approximation at the end of the cycle y_l can be represented as $y_l = \mathcal{M}_{l,C}(w_l)$ where $\mathcal{M}_{l,C}$ is a nonlinear function, since a Krylov subspace method (namely preconditioned GMRES(m_β)) is used as an approximate solver on the coarsest grid Ω_1. The

[1] In [99] the authors have introduced the complex shifted Laplace with a negative imaginary part for the shift in the case of first- or second-order radiation boundary conditions. Due to the PML formulation considered in this paper, we have used a shift with positive imaginary part to derive an efficient preconditioner as explained in [200, Section 3.3.2].

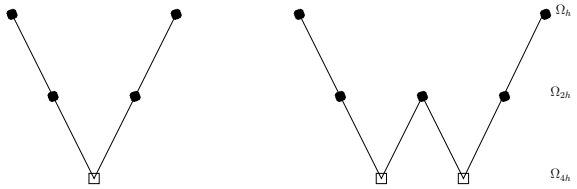


Figure 2.1.: Complex shifted Laplace multigrid cycle applied to $S_l^{(\beta)} y_l = w_l$ sketched in Algorithm 2.1. $\mathcal{M}_{3,V}$ (left) and of $\mathcal{M}_{3,F}$ (right). A $\bullet$ symbol represents a smoothing step, while the $\square$ symbol represents an approximate coarse grid solution.

multigrid cycle of Algorithm 2.1 is based on a Jacobi smoother as promoted in [98] and slightly differs from the original algorithm proposed in [98]. Indeed Erlangga et al. [98] have used the matrix-dependent interpolation operator of [269], a Galerkin coarse grid approximation to deduce the discrete coarse operators and an exact solution on the coarsest grid. For three-dimensional applications, Erlangga [96] and Riyanti et al. [207] have proposed a multigrid method with a two-dimensional semi-coarsening strategy combined with line-wise damped Jacobi smoothing in the third direction. A cycle of multigrid acting on this complex shifted Laplace operator is then considered as a preconditioner for the Helmholtz operator.

2.4.2. Properties

Since its introduction, this preconditioning technique based on a different partial differential equation has been widely used, see, e.g., [37, 65, 94, 93, 207, 233, 255] for applications in three dimensions. The theoretical properties of this preconditioner have been investigated in [123], where it has been shown that the eigenvalues of the preconditioned operator move to zero as the frequency increases. An immediate consequence is that a strong increase in terms of preconditioner applications is observed for the medium to high frequency range. This has been shown in [100] by Fourier Analysis in a one-dimensional setting. We note that a theoretical analysis in the framework of finite element discretization has been proposed more recently; see [114]. An illustration is given in Figure 2.2 for a realistic application described in Section 2.8. Although the frequency (10 Hz) is moderate, we observe Ritz and harmonic Ritz values close to the origin in the complex plane as expected. This induces an increase in terms of preconditioner applications as reported later in Section 2.8.

Hence, deflation or augmentation methods have to be employed in combination with preconditioning to improve the convergence rate of the numerical method; see Chapter 4. This has been recently pursued in [222]. Nevertheless, these techniques may be expensive especially for applications in three dimensions at high frequency.

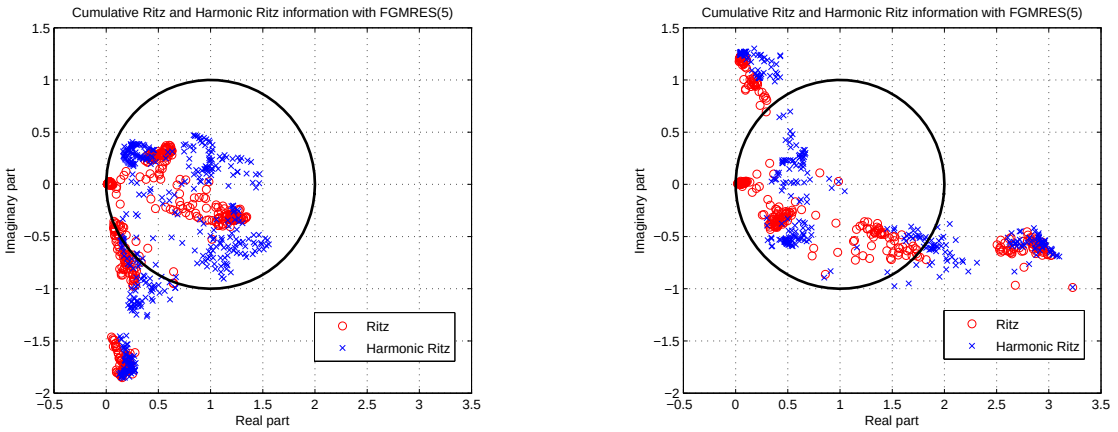


Figure 2.2.: EAGE/SEG Salt dome problem (f = 10 Hz, $927 \times 927 \times 287$ grid). Ritz and harmonic Ritz values (circles and crosses, respectively) of FGMRES(5) with two different variable preconditioners: $\mathcal{M}_{3,V}$ (left part) and $\mathcal{M}_{3,F}$ (right part) along convergence. Figure 7 of [58].

2.5. Basic two-grid preconditioner

We briefly examine the two-grid preconditioner proposed by Xavier Pinel in his PhD thesis [200]. We will use this geometric preconditioner in the numerical experiments related to the solution of linear systems with multiple right-hand sides presented in Chapter 4.

2.5.1. Algorithm and components

The two-grid cycle to be used as a preconditioner is sketched in Algorithm 2.2, where it is assumed that the initial approximation z_h^0 is equal to zero on Ω_h, denoted later by 0_h. As in [92, 250], polynomial smoothers based on GMRES [218] have been selected for both pre- and post-smoothing phases. Here a cycle of preconditioned GMRES(m_s) on Ω_h involves m_s matrix-vector products with A_h and $m_s\nu$ iterations of damped Jacobi. Finally, we note that the approximation at the end of the cycle z_h can be represented as $z_h = \mathcal{T}(v_h)$ where $\mathcal{T}$ is a nonlinear function due both to the use of a polynomial method based on GMRES as a smoother and to the approximate solution obtained on the coarse grid.

2.5.2. Properties

In the framework of indefinite Helmholtz problems with homogeneous velocity field, solving only approximately the coarse level problem has been analysed by rigorous Fourier analysis in [200]. Theoretical developments supported by numerical experiments have notably shown that solving the coarse level problem approximately may lead to an

Algorithm 2.2 Two-grid cycle applied to $A_h z_h = v_h$. $z_h = \mathcal{T}(v_h)$.

Input: Assume that the following is given

- $A_h \in \mathbb{C}^{n_h \times n_h}$ ▷ Helmholtz operator discretized on the fine grid Ω_h
- $A_H \in \mathbb{C}^{n_H \times n_H}$ ▷ Helmholtz operator discretized on the coarse grid Ω_H
- $v_h \in \mathbb{C}^{n_h}$ ▷ right-hand side given on Ω_h
- $z_h \in \mathbb{C}^{n_h}$ ▷ initial guess given on Ω_h

1: Polynomial pre-smoothing: Apply ϑ cycles of GMRES(m_s) to $A_h z_h = v_h$ with ν iterations of ω_h-Jacobi as a right preconditioner to obtain the approximation z_h^{ϑ}.
2: Restrict the fine level residual: $v_H = I_h^H (v_h - A_h z_h^{\vartheta})$.
3: Solve approximately the coarse problem $A_H z_H = v_H$ with initial approximation $z_H^0 = 0_H$: Apply ϑ_c cycles of GMRES(m_c) to $A_H z_H = v_H$ with ν_c iterations of ω_H-Jacobi as a right preconditioner to obtain the approximation z_H.
4: Perform the coarse level correction: $\widetilde{z_h} = z_h^{\vartheta} + I_H^h \, z_H$.
5: Polynomial post-smoothing: Apply ϑ cycles of GMRES(m_s) to $A_h z_h = v_h$ with initial approximation $\widetilde{z_h}$ and ν iterations of ω_h-Jacobi as a right preconditioner to obtain the final approximation z_h.

efficient two-grid preconditioner. We refer the reader to [200, Section 3.4] for a complete analysis on three-dimensional academic model problems. Nevertheless, one of the main difficulties related to the two-grid preconditioner presented in this section is that the coarse linear system becomes indefinite as the frequency grows due to the condition (2.5). This is illustrated in Figure 2.3. Consequently, to derive an efficient numerical method in the high frequency range, it is mandatory to find an efficient coarse grid preconditioner. This is the main goal of the new multigrid preconditioner presented next.

2.6. Improved two-grid preconditioner

We now introduce the main contribution of this chapter. This two-grid preconditioner will be later analysed in Section 2.7 and related detailed numerical experiments will be reported in Section 2.8.

2.6.1. Algorithm and components

We introduce a multigrid cycle acting on a complex shifted Laplace operator as a preconditioner for the coarse grid system $A_H z_H = v_H$ defined on Ω_H. The complex shifted Laplace operator is simply obtained by direct coarse grid discretization of equations (2.7,2.8,2.9) on Ω_H. The new cycle can be seen as a combination of two cycles defined on two different hierarchies. Firstly, a two-grid cycle using Ω_h and Ω_H only as fine and coarse levels respectively is applied to the Helmholtz operator. Secondly, a sequence of grids Ω_k ($k = 1, \cdots, l$) with the finest grid Ω_l defined as $\Omega_l := \Omega_H$ is introduced. On this second hierarchy a multigrid cycle applied to a complex shifted Laplace opera-

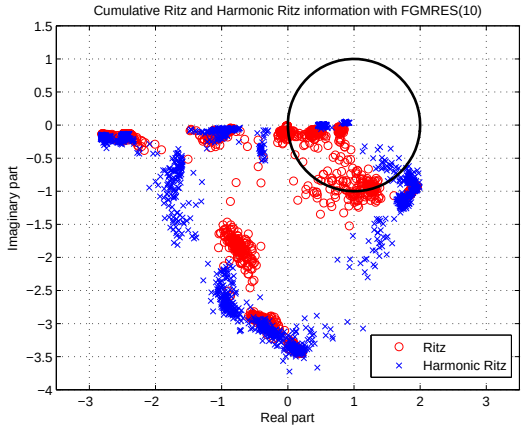


Figure 2.3.: EAGE/SEG Salt dome problem ($f = 10$ Hz, $927 \times 927 \times 287$ grid) using a basic two-grid preconditioner. Cumulative Ritz and harmonic Ritz values (circles and crosses, respectively) obtained when solving the coarse grid problems $A_H z_H = v_H$. FGMRES(10) is used as a Krylov subspace solver on the coarse level. Hence 10 Ritz (or harmonic) approximations are obtained per cycle.

tor $S_H^{(\beta)} := S_l^{(\beta)}$ is then used as a preconditioner when solving the coarse level system $A_H z_H = v_H$ of the two-grid cycle. The new combined cycle is sketched in Algorithm 2.3.

The notation $\mathcal{T}_{l,C}$ uses subscripts related to the cycle applied to the shifted Laplace operator (i.e. number of grids l of the second hierarchy and cycling strategy C (which can be of V, F or W type), respectively). The combined cycle then involves discretization of operators on $l+1$ grids in total. Hence later in the numerical experiments we will compare $\mathcal{T}_{l,C}$ with $\mathcal{M}_{l+1,C}$. Figure 2.4 shows a possible configuration with a three-grid cycle applied to the shifted Laplace operator. The combined cycle is related to the recursively defined K-cycle introduced in [193]. Nevertheless, we note that the combined cycle relies on a preconditioning operator on the coarse level that is different from the original operator. The approximation at the end of the cycle z_h can be represented as $z_h = \mathcal{T}_{l,C}(v_h)$ where $\mathcal{T}_{l,C}$ is a nonlinear function obtained as a combination of functions introduced in Sections 2.4 and 2.5, respectively. Consequently, this cycle leads to a variable nonlinear preconditioner which must be combined with an outer *flexible* Krylov subspace method [223, 224] and [253, Chapter 10].

2.6.2. Properties

As an illustration, Figure 2.5 represents the cumulative Ritz and harmonic Ritz information obtained along convergence on the coarse level. The approximations are located on the right part of the complex plane and are relatively clustered. This is a favourable situation when using flexible GMRES as an approximate coarse solver. The comparison

Algorithm 2.3 Combined cycle applied to $A_h z_h = v_h$. $z_h = \mathcal{T}_{l,C}(v_h)$.

Input: Assume that the following is given

- $A_h \in \mathbb{C}^{n_h \times n_h}$ ▷ Helmholtz operator discretized on the fine grid Ω_h
- $A_H \in \mathbb{C}^{n_H \times n_H}$ ▷ Helmholtz operator discretized on the coarse grid Ω_H
- $v_h \in \mathbb{C}^{n_h}$ ▷ right-hand side given on Ω_h
- $z_h \in \mathbb{C}^{n_h}$ ▷ initial guess given on Ω_h
- $S_k^{(\beta)} \in \mathbb{C}^{n_k \times n_k}$ $(k = 1, \cdots, l)$ ▷ complex shifted Laplace operators discretized on Ω_k

1: Polynomial pre-smoothing: Apply ϑ cycles of GMRES(m_s) to $A_h z_h = v_h$ with ν iterations of ω_h-Jacobi as a right preconditioner to obtain the approximation z_h^{ϑ}.
2: Restrict the fine level residual: $v_H = I_h^H (v_h - A_h z_h^{\vartheta})$.
3: Solve approximately the coarse problem $A_H z_H = v_H$ with initial approximation $z_H^0 = 0_H$: Apply ϑ_c cycles of FGMRES(m_c) to $A_H z_H = v_H$ preconditioned by a cycle of multigrid applied to $S_l^{(\beta)} y_l = w_l$ on $\Omega_l \equiv \Omega_H$ yielding $y_l = \mathcal{M}_{l,C}(w_l)$ to obtain the approximation z_H.
4: Perform the coarse level correction: $\widetilde{z_h} = z_h^{\vartheta} + I_H^h z_H$.
5: Polynomial post-smoothing: Apply ϑ cycles of GMRES(m_s) to $A_h z_h = v_h$ with initial approximation $\widetilde{z_h}$ and ν iterations of ω_h-Jacobi as a right preconditioner to obtain the final approximation z_h.

with Figure 2.3 is striking. We refer the reader to Section 2.8 for a complete analysis of the efficiency of the new multigrid preconditioner on a realistic three-dimensional problem.

We will investigate the properties of the improved two-grid preconditioner theoretically. We will mostly rely on rigorous Fourier analysis to select appropriate smoothers and to analyse the two-grid iteration error matrix. This is examined next.

2.7. Fourier analysis of multigrid preconditioners

In this section, we provide a two-grid rigorous Fourier analysis to select appropriate relaxation parameters in the smoother and to understand the convergence properties of the two-grid methods used as a preconditioner introduced in Sections 2.4, 2.5 and 2.6. For this analysis only, we consider a two-grid method based on a Jacobi smoother, standard coarsening, full-weighting, trilinear interpolation and exact solution on the coarse grid, applied to a model problem of Helmholtz type discretized with a standard second-order finite difference scheme. We refer the reader to [234, 239] for the theoretical foundations of rigorous Fourier analysis.

2.7.1. Notation specific to Fourier analysis

Throughout Section 2.7, we consider the complex shifted Laplace equation with a uniform wavenumber given by $k = 2\pi f/c$ on the unit cube $\Omega = [0,1]^3$ and with homogeneous

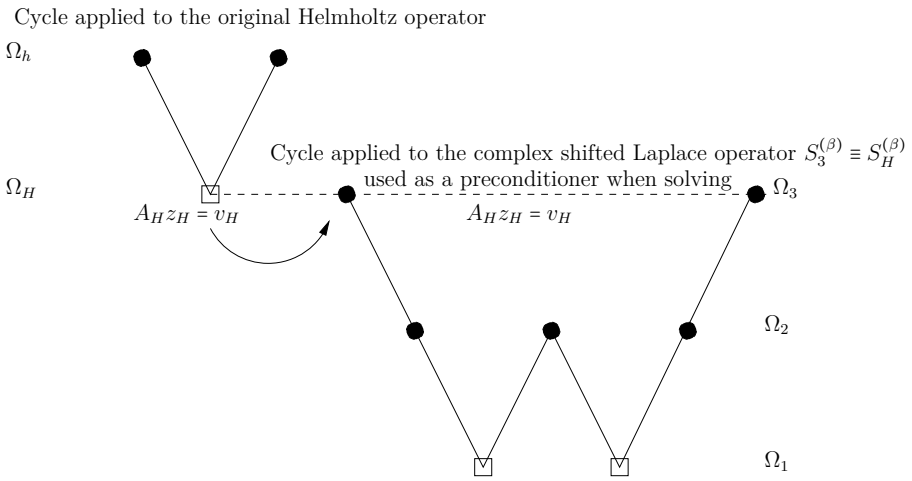


Figure 2.4.: Combined cycle applied to $A_h z_h = v_h$ sketched in Algorithm 2.3 using $\mathcal{T}_{3,F}$. The two-grid cycle is applied to the Helmholtz operator (left part), whereas the three-grid cycle to be used as a preconditioner when solving the coarse grid problem $A_H z_H = v_H$ is shown on the right part. This second multigrid cycle acts on the shifted Laplace operator with β as a shift parameter. Figure 1 of [58].

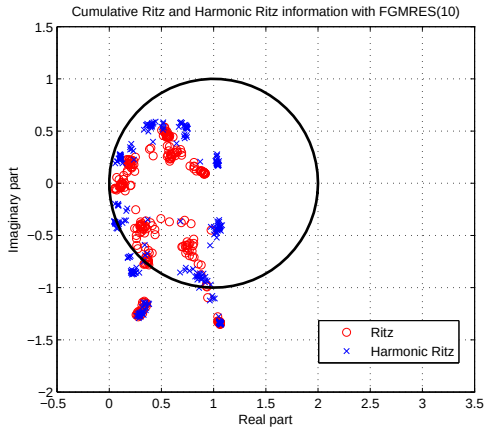


Figure 2.5.: EAGE/SEG Salt dome problem ($f = 10$ Hz, $927 \times 927 \times 287$ grid) using an improved two-grid preconditioner $\mathcal{T}_{2,V}$. Cumulative Ritz and harmonic Ritz values (circles and crosses, respectively) obtained when solving the coarse grid problems $A_H z_H = v_H$. FGMRES(10) is used as a Krylov subspace solver on the coarse level. Hence 10 Ritz (or harmonic) approximations are obtained per cycle.

Dirichlet boundary conditions on the boundary of the domain:

$$-\Delta u - \kappa_\beta^2 u = g \quad \text{in} \quad \Omega, \tag{2.10}$$
$$u = 0 \quad \text{on} \quad \partial\Omega, \tag{2.11}$$

with κ_β defined as $\kappa_\beta^2 = (1+\beta i)k^2$, where β denotes a real parameter lying in $[0,1]$. A classical tool in multigrid theory to deduce information about the two-grid convergence rate is based on a rigorous Fourier analysis (RFA) [247, Section 3.3.4]. To perform this analysis, we introduce some additional notation. Firstly, we discretize the model problem (2.10, 2.11) on an uniform mesh of step size $\chi = 1/n_\chi$. We denote by $L_\chi^{(\beta)}$ the corresponding discrete operator on the fine grid $\Omega_\chi = G_\chi \cap [0,1]^3$ where G_χ is the infinite grid and by $D_\chi^{(\beta)}$ the matrix corresponding to the diagonal part of $L_\chi^{(\beta)}$. The discrete eigenfunctions of $L_\chi^{(\beta)}$:

$$\varphi_\chi^{l_1,l_2,l_3}(x,y,z) = \sin(l_1\pi x)\,\sin(l_2\pi y)\,\sin(l_3\pi z)$$
$$\text{with } l_1,l_2,l_3 = 1,...,n_\chi - 1 \text{ and } (x,y,z) \in \Omega_\chi,$$

generate the space of all fine grid functions, $F(\Omega_\chi)$, and are orthogonal with respect to the discrete inner product on Ω_χ:

$$(v_\chi, w_\chi) := \chi^3 \sum_{(x,y,z)\in\Omega_\chi} v_\chi(x,y,z) w_\chi(x,y,z) \quad \text{with } v_\chi, w_\chi \in F(\Omega_\chi).$$

The space of all fine grid real-valued functions $F(\Omega_\chi)$ can be divided into a direct sum of (at most) eight-dimensional subspaces - called the 2χ-harmonics [247, Equation (3.4.1)] - :

$$\begin{aligned} E_\chi^{l_1,l_2,l_3} = \; & span[\varphi_\chi^{l_1,l_2,l_3}, -\varphi_\chi^{n_\chi-l_1,n_\chi-l_2,n_\chi-l_3}, -\varphi_\chi^{n_\chi-l_1,l_2,l_3}, \varphi_\chi^{l_1,n_\chi-l_2,n_\chi-l_3}, \\ & -\varphi_\chi^{l_1,n_\chi-l_2,l_3}, \varphi_\chi^{n_\chi-l_1,l_2,n_\chi-l_3}, -\varphi_\chi^{l_1,l_2,n_\chi-l_3}, \varphi_\chi^{n_\chi-l_1,n_\chi-l_2,l_3}], \\ & \text{for } l_1,l_2,l_3 = 1,\cdots,n_\chi/2. \end{aligned}$$

The dimension of $E_\chi^{l_1,l_2,l_3}$, denoted by $\eta_\chi^{l_1,l_2,l_3}$, is eight, four, two and one if zero, one, two or three of the indices l_1,l_2,l_3 is equal to $n_\chi/2$, respectively. Similarly as on the fine grid Ω_χ, we introduce the discrete eigenfunctions of the coarse grid operator $L_{2\chi}^{(\beta)}$ on the space of all coarse grid functions $F(\Omega_{2\chi})$ with $\Omega_{2\chi} = G_{2\chi} \cap [0,1]^3$:

$$\varphi_{2\chi}^{l_1,l_2,l_3}(x,y,z) = \sin(l_1\pi x)\,\sin(l_2\pi y)\,\sin(l_3\pi z),$$
$$\text{with } l_1,l_2,l_3 = 1,...,\frac{n_\chi}{2} - 1 \text{ and } (x,y,z) \in \Omega_{2\chi}.$$

$E_{2\chi}^{l_1,l_2,l_3}$ is then defined as $span[\varphi_{2\chi}^{l_1,l_2,l_3}]$ since the eigenfunctions of $L_{2\chi}$ coincide up to their sign on $\Omega_{2\chi}$ for $l_1,l_2,l_3 = 1,\cdots,n_\chi/2$ [247]. We denote later by ℓ the multi-index $\ell = (l_1,l_2,l_3)$, by $\mathcal{L}_\chi = \{\ell \mid 1 \le \max(l_1,l_2,l_3) < n_\chi/2\}$ and by $\mathcal{H}_\chi = \{\ell \mid n_\chi/2 \le \max(l_1,l_2,l_3) < n_\chi\}$ the sets of multi-indices corresponding to the low-frequency and high-frequency

harmonics, respectively. We also define the set $\mathcal{L}_\chi^= = \{\ell \,|1 \le \max(l_1, l_2, l_3) \le n_\chi/2\}$. Later in this section, the Fourier representation of a given discrete operator M_χ is denoted by $\widehat{M}_\chi$ and the restriction of $\widehat{M}_\chi$ to E_χ^ℓ with $\ell \in \mathcal{L}_\chi$ is denoted by $\widehat{M}_\chi(\ell) = \widehat{M}_{\chi|E_\chi^\ell}$ in short. The Fourier representation of the discrete Helmholtz operator $L_\chi^{(\beta)}$ and the Jacobi iteration matrix $J_\chi^{(\beta)}$ are denoted by $\widehat{L}_\chi^{(\beta)}$ and $\widehat{J}_\chi^{(\beta)}$, respectively. To write the Fourier representation of these operators in a compact form, we also introduce the ξ_i parameters such that $\xi_i = \sin^2\left(\frac{l_i\pi\chi}{2}\right)$ for $i = 1, 2, 3$. Finally we denote by $\chi = h$ the finest mesh grid size considered, n_h the corresponding number of points per direction and k_χ the wavenumber on the grid with mesh size χ.

2.7.2. Smoother analysis

The multigrid method acting on a complex shifted Laplace operator presented in Algorithm 2.1 is based on a Jacobi smoother as used in [98] in two dimensions. Indeed in [98] it has been numerically shown that this method enjoys good smoothing properties on all the grids of the hierarchy when the relaxation parameters ω_χ are well chosen. In Proposition 2.1, we give the Fourier representation of the Jacobi iteration matrix $J_\chi^{(\beta)}$ applied to the complex shifted Laplace matrix $L_\chi^{(\beta)}$. Then we derive related smoothing factors and, by numerical experiments, we deduce appropriate damping parameters to obtain good smoothing properties in three dimensions.

Proposition 2.1. *The harmonic spaces E_χ^ℓ for $\ell \in \mathcal{L}_\chi^=$ are invariant under the Jacobi iteration matrix $J_\chi^{(\beta)} = I_\chi - \omega_\chi (D_\chi^{(\beta)})^{-1} L_\chi^{(\beta)}$ ($J_\chi^{(\beta)} : E_\chi^\ell \longrightarrow E_\chi^\ell$, for $\ell \in \mathcal{L}_\chi^=$). The operator $J_\chi^{(\beta)}$ is orthogonally equivalent to a block diagonal matrix of (at most) 8×8 blocks defined as:*

$$\widehat{J}_\chi^{(\beta)}(\ell) \;=\; I_{\eta_\chi^\ell} - \left(\frac{\omega_\chi \chi^2}{6 - (\kappa_\beta \chi)^2}\right) \widehat{L}_\chi^{(\beta)}(\ell), \quad \ell \in \mathcal{L}_\chi^=, \tag{2.12}$$

where $\widehat{L}_\chi^{(\beta)}$ denotes the representation of the complex shifted Laplace operator $L_\chi^{(\beta)}$ with respect to the space E_χ^ℓ and η_χ^ℓ the dimension of E_χ^ℓ, respectively. With the notation introduced in Section 2.7.1, if $\ell \in \mathcal{L}_\chi$, the representation of $\widehat{L}_\chi^{(\beta)}$ with respect to E_χ^ℓ is a diagonal matrix defined as:

$$\widehat{L}_\chi^{(\beta)}(\ell) \;=\; diag\left(\frac{4}{\chi^2}\begin{pmatrix} (\xi_1+\xi_2+\xi_3) \\ (3-\xi_1-\xi_2-\xi_3) \\ (1-\xi_1+\xi_2+\xi_3) \\ (2+\xi_1-\xi_2-\xi_3) \\ (1+\xi_1-\xi_2+\xi_3) \\ (2-\xi_1+\xi_2-\xi_3) \\ (1+\xi_1+\xi_2-\xi_3) \\ (2-\xi_1-\xi_2+\xi_3) \end{pmatrix} \begin{matrix} -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \end{matrix}\right), \quad \ell \in \mathcal{L}_\chi. \tag{2.13}$$

If one of the indices of ℓ equals $n_\chi/2$, $\widehat{L}_\chi^{(\beta)}(\ell)$ degenerates to a diagonal matrix of dimension η_χ^ℓ. Its entries then correspond to the first η_χ^ℓ entries of the matrix given on the right-hand side of relation (2.13).

Proof. See Proposition 1 in [58]. Obviously, since the eigenfunctions spanning E_χ^ℓ are eigenfunctions of $L_\chi^{(\beta)}$, the harmonic spaces E_χ^ℓ ($\ell \in \mathcal{L}_\chi$) are invariant under $L_\chi^{(\beta)}$ and hence invariant under $J_\chi^{(\beta)}$. The representation of $L_\chi^{(\beta)}$ with respect to the harmonic space E_χ^ℓ is obtained by writing the eigenvalues of the basis functions of E_χ^ℓ in terms of ξ_i, a straightforward calculation that only involves trigonometric identities. □

The representation of the Jacobi iteration matrix in the Fourier space obtained in Proposition 2.1 allows us to easily investigate its smoothing properties, i.e., to compute the smoothing factor μ versus various parameters (β, mesh grid size χ, wavenumber k_χ and relaxation parameter ω_χ, respectively). With ν denoting the number of relaxation sweeps, the smoothing factor $\mu(\beta, \chi, k_\chi, \omega_\chi)$ is defined as follows [265]

$$\mu(\beta, \chi, k_\chi, \omega_\chi) \quad = \quad \max_{\ell \in \mathcal{L}_\chi^=} |(\rho(\widehat{Q}_\chi(\ell)\ (\widehat{J}_\chi^{(\beta)}(\ell))^\nu)^{1/\nu}|, \tag{2.14}$$

where $\widehat{Q}_\chi$ is the matrix representation of a projection operator that annihilates the low-frequency error components and leaves the high-frequency components unchanged [234], e.g., $\widehat{Q}_\chi(\ell) = diag((0,1,1,1,1,1,1,1)^T)$ for $\ell \in \mathcal{L}_\chi$. In addition if we assume that $\kappa_\beta \chi$ (or similarly $k_\chi \chi$) is a given constant (which is the case in practice due to the stability condition to be satisfied) it is then possible to deduce the supremum $\mu^\star(\beta, \chi, k_\chi, \omega_\chi)$ of the smoothing factor over χ as

$$\mu^\star(\beta, \chi, k_\chi, , \omega_\chi) \quad = \quad \max\left\{ |1 - \omega_\chi \frac{2 - \kappa_\beta^2 \chi^2}{6 - \kappa_\beta^2 \chi^2}|, |1 - \omega_\chi \frac{12 - \kappa_\beta^2 \chi^2}{6 - \kappa_\beta^2 \chi^2}| \right\}, \tag{2.15}$$

or similarly

$$\mu^\star(\beta, \chi, k_\chi, \omega_\chi) = \max\left\{ |1 - \omega_\chi + \frac{4\omega_\chi}{6 - (1 + i\beta) k_\chi^2 \chi^2}|, |1 - \omega_\chi - \frac{6\omega_\chi}{6 - (1 + i\beta) k_\chi^2 \chi^2}| \right\}. \tag{2.16}$$

For a fixed value of $k_\chi \chi$ this formula can then give guidance in choosing the optimal relaxation parameters and in understanding how the optimal value of the relaxation parameter $\omega_\chi^\star$ depends on $k_\chi \chi$ and on β, respectively. Indeed a simple calculation gives the real-valued optimal relaxation parameter as

$$\omega_\chi^\star = 1 - \frac{1}{7 - k_\chi^2 \chi^2}.$$

We notice that the optimal value of the relaxation parameter does not depend on the shift parameter β and note that we recover the optimal relaxation parameter and the supremum of the smoothing factor of the Jacobi method for the Poisson equation in three dimensions when k_χ is set to zero [247, Section 2.9.2].

2.7.3. Preconditioner analysis

Assumptions on the components of the cycle

In this paragraph, we assume that both the fine grid operator and the smoother leave the spaces E_χ^ℓ invariant for $\ell \in \mathcal{L}_\chi^=$. As shown in Proposition 2.1, $L_\chi^{(\beta)}$ and the corresponding Jacobi iteration matrix $J_\chi^{(\beta)}$ do satisfy this invariance property. Furthermore we assume that the transfer operators $I_\chi^{2\chi}$, $I_{2\chi}^{\chi}$ satisfy the following relations

$$I_\chi^{2\chi} : E_\chi^\ell \to span[\varphi_{2\chi}^\ell], \quad I_{2\chi}^{\chi} : span[\varphi_{2\chi}^\ell] \to E_\chi^\ell, \quad \text{for } \ell \in \mathcal{L}_\chi. \tag{2.17}$$

and that the coarse discretization operator leaves the subspace $span[\varphi_{2\chi}^\ell]$ invariant for $\ell \in \mathcal{L}_\chi$. We note that the discrete coarse Helmholtz matrix $L_{2\chi}^{(\beta)}$ satisfies this last property and that the trilinear interpolation and its adjoint also satisfy relation (2.17) [247].

Proposition 2.2. *If the previous assumptions are satisfied, the iteration matrix of the two-grid cycle ($M_\chi^{(\beta)} : E_\chi^\ell \longrightarrow E_\chi^\ell$, for $\ell \in \mathcal{L}_\chi^=$) leaves the spaces of 2χ-harmonics E_χ^ℓ with an arbitrary $\ell \in \mathcal{L}_\chi^=$ invariant. The Fourier representation of the two-grid iteration matrix $M_\chi^{(\beta)}$ is as a block-diagonal matrix of (at most) 8×8 blocks defined as:*

$$\widehat{M}_\chi^{(\beta)}(\ell) \;=\; (\widehat{J}_\chi^{(\beta)}(\ell))^\nu \widehat{K}_{\chi,2\chi}^{(\beta)}(\ell)(\widehat{J}_\chi^{(\beta)}(\ell))^\nu \quad \textit{for} \quad \ell \in \mathcal{L}_\chi^=, \tag{2.18}$$

with $\widehat{K}_{\chi,2\chi}^{(\beta)}(\ell) = I_8 - [c\,d^T]/\Lambda_{2\chi}^{(\beta)}$ *if* $\ell \in \mathcal{L}_\chi$, *where* $\Lambda_{2\chi}^{(\beta)} = \dfrac{4}{\chi^2}((1-\xi_1)\xi_1 + (1-\xi_2)\xi_2 + (1-\xi_3)\xi_3) - \kappa_\beta^2$ *and* $c \in \mathbb{R}^8$, $d \in \mathbb{C}^8$, *are defined as follows*

$$\begin{cases} c_1 = (1-\xi_1)(1-\xi_2)(1-\xi_3),\ c_2 = \xi_1\xi_2\xi_3,\ c_3 = \xi_1(1-\xi_2)(1-\xi_3),\ c_4 = (1-\xi_1)\xi_2\xi_3, \\ c_5 = (1-\xi_1)\xi_2(1-\xi_3),\ c_6 = \xi_1(1-\xi_2)\xi_3,\ c_7 = (1-\xi_1)(1-\xi_2)\xi_3,\ c_8 = \xi_1\xi_2(1-\xi_3), \\ d = \widehat{L}_\chi^{(\beta)}(\ell)\ c,\ \textit{where } \widehat{L}_\chi^{(\beta)}(\ell) \textit{ is defined in equation (2.13).} \end{cases}$$

If one of the indices of ℓ is equal to $n_\chi/2$, $\widehat{K}_{\chi,2\chi}^{(\beta)}(\ell)$ is reduced to the identity matrix of dimension η_χ^ℓ.

Proof. See Proposition 2 in [58]. Under the assumptions given above, it is straightforward to prove that the iteration matrix of the two-grid cycle leaves E_χ^ℓ for $\ell \in \mathcal{L}_\chi^=$ invariant. We obtain formula (2.18) by just combining the Fourier representation of each of its components. The complete details of these trigonometric calculations can be found in [200, Section 3.3.1]. □

For the sake of brevity, the reader can find in [58] the complete numerical results related to the rigorous Fourier analysis. This analysis has allowed us to select appropriate relaxation parameters in the Jacobi method that lead to acceptable smoothing factors on all the grids of a complex shifted multigrid method in three dimensions (Figure 3

of [58]). We have also shown the suitability of the complex shifted multigrid preconditioner on the coarse level of a combined two-grid method (left part of Figure 4 of [58]). Although rigorous Fourier analysis corresponds to a simplified analysis, numerical experiments performed on a homogeneous velocity field studied in Section 5.2 of [58] have supported these conclusions. We next investigate the performance of the preconditioner on a realistic three-dimensional heterogeneous Helmholtz problem.

2.8. Numerical results on the SEG/EAGE Salt dome model

We illustrate the performance of the various preconditioners presented in Sections 2.4, 2.5 and 2.6 combined with Flexible GMRES(m) for the solution of the acoustic Helmholtz problem (2.2, 2.3, 2.4) on a realistic heterogeneous velocity model used as a benchmark in the geophysics community. The SEG/EAGE Salt dome model [11] is a velocity field containing a salt dome in a sedimentary embankment. It is defined in a parallelepiped domain of size $13.5 \times 13.5 \times 4.2\ km^3$. The minimum value of the velocity is 1500 $m.s^{-1}$ and its maximum value is 4481 $m.s^{-1}$. This test case is considered as challenging due to both the occurrence of a geometrically complex structure (salt dome) and to the truly large dimensions of the computational domain. In the subsections 2.8.2 to 2.8.4, we address three important aspects: the robustness with respect to frequency, the strong scalability properties and the complexity analysis of the numerical method.

2.8.1. Settings

In the two-grid cycle of Algorithm 2.2, we consider as a smoother the case of one cycle of GMRES(2) preconditioned by two iterations of damped Jacobi ($\vartheta = 1$, $m_s = 2$ and $\nu = 2$), a restarting parameter equal to $m_c = 10$ for the preconditioned GMRES method used on the coarse level and a maximum number of coarse cycles equal to $\vartheta_c = 10$. In the complex shifted multigrid cycle of Algorithm 2.1, we use a shift parameter equal to $\beta = 0.5$ and two iterations of damped Jacobi as a smoother ($\nu_\beta = 2$). On the coarsest level we consider as an approximate solver one cycle of GMRES(10) preconditioned by two iterations of damped Jacobi ($\vartheta_\beta = 1$, $m_\beta = 10$ and $\nu_\beta = 2$). The previous parameters were also used in Algorithm 2.3, with an exception made for ϑ_c set to 2. Finally, the relaxation coefficients considered in the Jacobi method were determined by rigorous Fourier analysis and are given by the following relation [58]

$$(\omega_h, \omega_{2h}, \omega_{4h}, \omega_{8h}) = (0.8, 0.8, 0.2, 1). \tag{2.19}$$

We consider a value of the restarting parameter of the outer Krylov subspace method equal to $m = 5$ as in [59, 200]. The unit source is located at

$$(s_1, s_2, s_3) = (h\ n_{x_1}/2, h\ n_{x_2}/2, h\ (n_{PML} + 1))$$

where, e.g., n_{x_1} denotes the number of points in the first direction. A zero initial guess x_h^0 is chosen and the iterative method is stopped when the Euclidean norm of the

residual normalized by the Euclidean norm of the right-hand side satisfies the following relation

$$\frac{\|b_h - A_h x_h\|_2}{\|b_h\|_2} \leq 10^{-5}. \tag{2.20}$$

2.8.2. Robustness with respect to the frequency

As pointed out earlier, an increase of the frequency does lead to an increase in the number of unknowns. Thus it is of paramount importance to analyse the behaviour of the numerical methods when the frequency grows. In this section we analyse the different preconditioners in this respect. The numerical results presented here (see [58]) have been obtained on Babel, a IBM Blue Gene/P computer located at IDRIS (each node of Babel is equipped with 4 PowerPC 450 cores at 850 Mhz) using a Fortran 90 implementation with MPI [134] in complex single precision arithmetic (see [247, Chapter 6] for the practical aspects related to the parallelization of geometric multigrid). Physical memory on a given node (4 cores) of Babel is limited to 2 GB. This code was compiled by the IBM compiler suite with the best optimization options and linked with the vendor BLAS and LAPACK subroutines.

$\mathcal{T}$						
f(Hz)	h	Grid	# Cores	Prec	T (s)	M (GB)
2.5	60	$231 \times 231 \times 71$	4	12	146	0.6
5	30	$463 \times 463 \times 143$	32	25	316	4.5
10	15	$927 \times 927 \times 287$	256	71	927	35.9
20	7.5	$1855 \times 1855 \times 575$	2048	248	3346	288.1
40	3.75	$3711 \times 3711 \times 1149$	16384	$1000^{\dagger}$	13912	2304.1
$\mathcal{T}_{2,V}$						
f(Hz)	h	Grid	# Cores	Prec	T (s)	M (GB)
2.5	60	$231 \times 231 \times 71$	4	11	98	0.6
5	30	$463 \times 463 \times 143$	32	16	147	4.6
10	15	$927 \times 927 \times 287$	256	28	270	36.6
20	7.5	$1855 \times 1855 \times 575$	2048	73	748	293.8
40	3.75	$3711 \times 3711 \times 1149$	16384	283	3101	2349.9
$\mathcal{M}_{3,V}$						
f(Hz)	h	Grid	# Cores	Prec	T (s)	M (GB)
2.5	60	$231 \times 231 \times 71$	4	98	132	0.5
5	30	$463 \times 463 \times 143$	32	217	300	3.8
10	15	$927 \times 927 \times 287$	256	445	638	30.5
20	7.5	$1855 \times 1855 \times 575$	2048	2485	4102	244.8
40	3.75	$3711 \times 3711 \times 1149$	16384	$8000^{\dagger}$	-	1957.8
$\mathcal{M}_{3,F}$						
f(Hz)	h	Grid	# Cores	Prec	T (s)	M (GB)
2.5	60	$231 \times 231 \times 71$	4	122	193	0.5
5	30	$463 \times 463 \times 143$	32	184	298	3.8
10	15	$927 \times 927 \times 287$	256	334	561	30.5
20	7.5	$1855 \times 1855 \times 575$	2048	2149	3764	244.8
40	3.75	$3711 \times 3711 \times 1149$	16384	$8000^{\dagger}$	-	1957.8

Table 2.1.: Robustness with respect to frequency. Preconditioned flexible methods for the solution of the Helmholtz equation for the heterogeneous velocity field EAGE/SEG Salt dome using a second-order discretization with 10 points per wavelength such that relation (2.5) is satisfied. Prec denotes the number of preconditioner applications, T the total computational time in seconds and M the requested memory in GB. Two-grid ($\mathcal{T}$), complex shifted multigrid cycles ($\mathcal{M}_{3,V}$, $\mathcal{M}_{3,F}$) and combined cycles ($\mathcal{T}_{2,V}$) are applied as a preconditioner of FGMRES(5). Numerical experiments performed on a IBM BG/P computer. A † superscript indicates that the maximum number of preconditioner applications has been reached. Table IV of [58].

Table 2.1 collects the number of preconditioner applications (Prec), computational

times (T) and maximum requested memory (M) for these variants (see also [58, Figure 8] for the plot of the convergence histories of the different numerical methods). With respect to the two-grid cycle $\mathcal{T}$, the new combined cycle $\mathcal{T}_{2,V}$ is found to require a reduced number of preconditioner applications. Indeed, if we consider the case of f = 20 Hz, we observe a significant reduction of preconditioner applications when comparing the two-grid preconditioner $\mathcal{T}$ with the combined two-grid cycle $\mathcal{T}_{2,V}$ (248 versus 73). This also leads to a dramatic reduction of computational times (3346 s versus 748 s at f = 20 Hz). The $\mathcal{T}_{2,V}$ strategy always delivers the minimum computational times among the four preconditioners with a clear advantage at medium to large frequencies. In this range the complex shifted Laplace multigrid preconditioner is found to require a large number of preconditioner applications, as expected. Nevertheless, we would like to stress that the shifted preconditioner presented in Algorithm 2.1 is based on a combination of standard multigrid components. It is most likely that the use of Galerkin coarse grid approximation or of operator-dependent transfer operators would be beneficial to improve the properties of the preconditioner when considering heterogeneous Helmholtz problems. Despite the simplicity of the shifted preconditioner we remark that both $\mathcal{M}_{3,V}$ and $\mathcal{M}_{3,F}$ strategies are more attractive than the two-grid preconditioner $\mathcal{T}$ in terms of computational times at small to medium range frequencies (2.5 Hz, 5 Hz and 10 Hz respectively). However at high frequencies (20 Hz and 40 Hz) a significant increase in terms of preconditioning applications is observed for both $\mathcal{M}_{3,V}$ and $\mathcal{M}_{3,F}$. We also notice that a shifted preconditioner based on a F-cycle is preferable when large frequencies are considered, i.e. solving approximately the coarse problem twice in a given cycle is found to be beneficial to the outer convergence.

2.8.3. Strong scalability

Hereafter, we only consider the $\mathcal{T}_{2,V}$ preconditioner which has been found to be efficient in Section 2.8.2 (see also [58] for additional numerical experiments that support this conclusion). The numerical results presented in Sections 2.8.3 and 2.8.4 have been obtained on Turing, a IBM Blue Gene/Q computer located at IDRIS (each node of Turing is equipped with 16 PowerPC A2-64 bit cores at 1.6 Ghz) using a Fortran 90 implementation with MPI in complex single precision arithmetic. Physical memory on a given node (16 cores) of Turing is limited to 16 GB.

$\mathcal{T}_{2,V}$					
f(Hz)	Grid	# Cores	Prec	T (s)	τ_s
20	$2303 \times 2303 \times 767$	16384	29	586	1.00
20	$2303 \times 2303 \times 767$	32768	29	302	0.97
20	$2303 \times 2303 \times 767$	65536	29	164	0.89
20	$2303 \times 2303 \times 767$	131072	29	87	0.84

Table 2.2.: Strong scalability analysis. Preconditioned flexible methods for the solution of the Helmholtz equation for the heterogeneous velocity field EAGE/SEG Salt dome using a dispersion minimizing discretization scheme with 10 points per wavelength such that relation (2.6) is satisfied. Prec denotes the number of preconditioner applications, T the total computational time in seconds and τ_s a scaled parallel efficiency defined in relation (2.21). $\mathcal{T}_{2,V}$ is applied as a preconditioner for FGMRES(5). Numerical experiments performed on a IBM BG/Q computer.

We are interested in the strong scalability properties of the numerical method. Hence, we consider the acoustic wave propagation problem at a fixed frequency (20 Hz) on a growing number of cores. In these numerical experiments, we consider the dispersion minimizing discretization scheme [143]; see Section 2.3.2. The step size h is determined by relation (2.6) with f_{ref} = 10 Hz, h_{ref} = 15 and q_{ref} = 4. Table 2.2 collects the number of preconditioner applications (Prec) and computational times (T) versus the number of cores. We note that the number of preconditioner applications is independent of the number of cores, which is a nice property. Next, we define a scaled parallel efficiency as

$$\tau_s = \frac{T_{ref}}{T} \Big/ \frac{Cores}{Cores_{ref}}, \tag{2.21}$$

where T_{ref} and $Cores_{ref}$ denote reference values for the computational time and number of cores, respectively. We collect the corresponding values in Table 2.2. We note that a perfect scaling corresponds to the value of 1. In practice, we note that τ_s is close to this value. Only the last numerical experiment performed on 131072 cores leads to a moderate degradation in terms of scaled parallel efficiency. This is partly due to the increased number of communications, which leads to a significant decrease in the ratio computation/communication.

2.8.4. Complexity analysis

$\mathcal{T}_{2,V}$					
f(Hz)	Grid	# Cores	Prec	T (s)	M (TB)
15	$1586 \times 1586 \times 492$	131072	19	30	0.56
20	$2303 \times 2303 \times 767$	131072	29	87	1.67
25	$3071 \times 3071 \times 1023$	131072	37	236	3.79
30	$3839 \times 3839 \times 1279$	131072	45	552	7.20
35	$4607 \times 4607 \times 1535$	131072	57	1158	12.2
40	$5631 \times 5631 \times 1791$	131072	69	2458	20.9

Table 2.3.: Complexity analysis. Preconditioned flexible methods for the solution of the Helmholtz equation for the heterogeneous velocity field EAGE/SEG Salt dome. A dispersion minimizing discretization scheme with 10 points per wavelength is used such that relation (2.6) is satisfied. Prec denotes the number of preconditioner applications, T the total computational time in seconds and M the requested memory in TB. $\mathcal{T}_{2,V}$ is applied as a preconditioner of FGMRES(5). Numerical experiments performed on a IBM BG/Q computer.

We analyse the complexity of the numerical method with respect to the frequency or to the problem size. In this experiment, the number of cores is kept fixed, while the frequency grows from 15 Hz to 40 Hz. The case of $f = 40$ Hz leads to a linear system with approximately 56.7 billion unknowns, the solution of which is certainly out of reach of numerical methods based on sparse direct methods described in Section 2.2. Results are given in Table 2.3. The number of preconditioner applications is rather moderate and is found to grow almost linearly with respect to the frequency. This linear dependency has been also observed for the complex shifted Laplace preconditioner in relation with other dispersion minimizing finite difference schemes [65], although on problems of smaller size. This behaviour is a quite satisfactory result, since huge linear systems can be solved in a reasonable amount of time on a parallel distributed memory machine. This result is especially interesting in the context of inverse problems (here acoustic full waveform inversion), where the solution of forward problems represents a major computational kernel. Figure 2.6 shows the evolution of the required memory (M) and computational time (T) versus the problem size (left part and right part, respectively). If N denotes the total number of unknowns, the computational time T is found to behave asymptotically as N^{α} with $\alpha = 1.32$. Finally, the memory requirements grow linearly with the problem size, as expected since no sparse factorization is involved either at the global or local levels in the multigrid preconditioner. This is also a nice

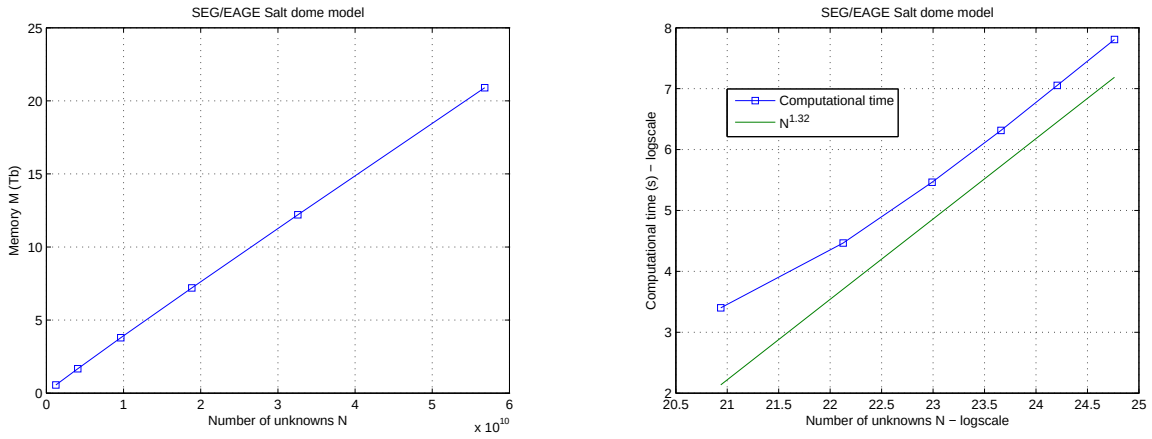


Figure 2.6.: Complexity analysis of the improved two-grid preconditioned Krylov subspace method. Evolution of memory requirements and computational time versus problem size. EAGE/SEG Salt dome using a dispersion minimizing discretization scheme with 10 points per wavelength such that relation (2.6) is satisfied. Results of Table 2.3.

feature of the numerical method.

2.9. Additional comments and conclusions

Summary In this chapter, we have focused on the analysis of multigrid preconditioners for the solution of wave propagation problems related to acoustic imaging. We have briefly reviewed the literature on multilevel preconditioners including domain decomposition and multigrid methods. As a main contribution, we have proposed a two-grid preconditioner for the numerical solution of Helmholtz problems in three-dimensional heterogeneous media. This two-grid cycle is applied directly to the original Helmholtz operator and relies on an approximate coarse grid solution. A second multigrid method applied to a complex shifted Laplace operator is then used as a preconditioner for the approximate solution of this coarse problem. Next, we have studied the convergence properties of this preconditioner with rigorous Fourier analysis. Finally, we have highlighted the efficiency of the new multigrid preconditioner on a concrete application in geophysics requiring the solution of indefinite problems of huge dimension (billions of unknowns) obtained after discretization by standard second order or dispersion minimizing finite difference schemes. Numerical results have demonstrated the usefulness of the combined algorithm on a realistic three-dimensional application at high frequency. Finally, a complexity analysis has been provided to close this chapter. This analysis has shown that the proposed numerical method is neither *scalable* with respect to the frequency nor *optimal*, which is unfortunately a common feature of most approaches for the solution of three-dimensional heterogeneous Helmholtz problems; however, see [168, 169, 270] for promising algorithms.

Collaboration The geometric multigrid preconditioner for heterogeneous Helmholtz problems has been developed in collaboration with Henri Calandra (TOTAL), Serge Gratton and Xavier Pinel with financial support from TOTAL. I have collaborated with Jean Piquet and Michel Visonneau on geometric multigrid methods for the applications in Computational Fluid Dynamics.

Software realization The proposed numerical methods have been written in Fortran 90 and MPI with linear algebra kernels relying on BLAS/LAPACK. During the project, the simulation code has been ported and validated on many parallel computing platforms (both internal and external). In this chapter, I have decided to report numerical experiments only on IBM BG/P or BG/Q platforms and refer the reader to [162, 200] for simulations on other computing platforms. Specific versions of the code have been integrated by Xavier Pinel and Rafael Lago, respectively, into the DIVA library at TOTAL. A pure MPI version of the code has been used on massively parallel platforms. A natural extension would be to consider a hybrid implementation based on MPI and OpenMP for addressing modern multicore architectures more efficiently.

Short-term prospectives In the wave propagation community, designing efficient and robust numerical methods for the solution of heterogeneous Helmholtz problems is a long standing question. In [58], we contributed to the design of geometric two-grid preconditioners and proposed numerical experiments in the context of seismic imaging. The proposed multigrid preconditioners have been recently used in the solution of inverse problems in seismics in the low frequency regime; see [77, 78]. A detailed comparison of the efficiency of the various numerical methods (including, e.g., advanced sparse direct solvers, domain decomposition and multigrid) on benchmark problems would be very helpful for the whole community. Finally, the simulation of acoustic-elastic wave propagation problems in exploration seismology becomes an emerging trend. As short-term perspectives, finding the appropriate approximation method and designing a related multigrid preconditioner in such a context is highly relevant. To conclude, to address the coming of extreme scale computers, we are aware that both the multilevel preconditioner and the outer Krylov subspace method need to be revised. Communication-avoiding or minimizing multigrid and Krylov subspace methods must be designed for such a purpose. This appears as an important milestone before addressing the numerical solution of both forward and inverse problems in seismic imaging on future computing platforms.

3. Non-overlapping domain decomposition methods for hp finite element methods

3.1. Objectives and contributions

3.1.1. Objectives

hp finite element methods are specific finite element methods, where the approximation in each element of maximum diameter h is a polynomial of degree k (see, e.g., [220] and references therein). Higher accuracy of the discrete solution is achieved by refining the mesh and/or by increasing the polynomial degree. hp finite element methods have been used in many applications in Computational Science and Engineering (CSE) such as Computational Fluid Dynamics, electromagnetics or structural mechanics; we refer the reader to the monographs [152, 220] for further details and references. This discretization method is known to be especially appropriate when accurate and/or dispersion free solutions of partial differential equations are required. Theoretically, a major feature is that hp finite element approximations converge exponentially fast when the mesh is refined using a suitable combination of h-refinements and p-refinements [17]. The exponential convergence makes the method a very attractive choice compared to most other finite element methods which only converge at an algebraic rate. When simulating physical phenomena exhibiting boundary layers or singularities, geometrically refined meshes towards corners, edges or faces must be used to guarantee this rate of convergence [220]. As a consequence, two- or three-dimensional meshes with high aspect ratios have to be employed in practice. Hence, the condition number of the stiffness matrix rapidly deteriorates: it grows exponentially with the spectral polynomial degree k [220]. The solution of such linear systems with iterative methods is thus especially difficult. In this chapter, we consider non-overlapping domain decomposition methods for the solution of partial differential equations discretized by hp finite element methods in two and three dimensions. Our objectives are twofold

- to propose and analyse popular non-overlapping domain decomposition methods used as preconditioners of Krylov subspace methods for the solution of linear systems resulting from the hp finite element discretization of scalar or vector problems,
- to provide detailed numerical experiments supporting the theoretical condition number bounds that have been established.

In this chapter, we mainly concentrate on the analysis of the domain decomposition preconditioners and provide condition number bounds of the preconditioned operator.

Our main interest is to investigate if we can design *scalable* domain decomposition preconditioners for the solution of elliptic partial differential equations discretized with the *hp* finite element method on anisotropic meshes in the sense of Definition 1.2. We point out that non-overlapping domain decomposition preconditioners are natural candidates for parallel preconditioners, see, e.g., [156, 226, 246] for related discussions. This important topic is not touched in this chapter however. Finally, we refer the reader to the monographs [152, 220, 227] for the various aspects related to both the analysis and the efficient implementation of *hp* finite element methods.

3.1.2. Contributions

The main contributions presented in this chapter are

- the analysis of two non-overlapping domain decomposition methods of balancing Neumann-Neumann [172] and one-level FETI type [103] on anisotropic meshes for the solution of scalar problems in two and three dimensions with the *hp* finite element method,
- the numerical illustration of the performance of the proposed domain decomposition preconditioners on selected academic examples.

The theoretical contribution extends the pioneering theoretical developments of Pavarino [198] concerning the design of non-overlapping domain decomposition preconditioners of Neumann-Neumann type for spectral element methods. Indeed, Pavarino has proved a condition number bound for the preconditioned operator P_{NN} in [198] as

$$\kappa(P_{NN}) \leq C\ (1 + log(k))^2,$$

where the constant C is independent not only of the polynomial degree k and of the number of substructures, but also of the values of the coefficients in the partial differential equation; see [246, Chapter 7] for further details on domain decomposition preconditioners for spectral finite element methods.

As additional contributions, we have also contributed to the theory of iterative substructuring methods for the solution of *hp* finite element approximations of *vector* problems in two dimensions in [244]. As an application, we have considered the solution of algebraic systems arising from edge element approximations in two dimensions [245]. Similarly to the scalar case, we have proposed an algorithm with a rate of convergence that is independent of possibly large jumps of the coefficients and mesh aspect ratios. We refer the reader to [245] for additional theory and numerical results supporting the theory. For the sake of brevity, these contributions are not presented here.

3.1.3. Synopsis

In Section 3.2, we review a few issues related to *hp* finite element approximations that appear as useful later. Then, we successively discuss the analysis of the balancing

Neumann-Neumann preconditioner in Section 3.3 and of the one-level FETI preconditioner in Section 3.4, respectively. In both sections, we derive the analysis of the preconditioners in matrix form to make easier the understanding of the related algorithms. Numerical experiments are provided in Sections 3.3.4 and 3.4.4, respectively. Finally, in Section 3.5, we provide the basis of a theory required to prove the condition number bounds given in Sections 3.3.2 and 3.4.2 in a variational setting. Conclusions are proposed in Section 3.6.

3.2. *hp* finite element approximation on geometrically refined meshes

We first introduce the academic problem that is considered throughout this chapter and briefly review a few aspects related to *hp* finite element approximations, meshes and domain partitioning, respectively.

3.2.1. Problem setting

We consider a linear, elliptic problem on a bounded polyhedral domain $\Omega \subset \mathbb{R}^3$ of unit diameter, formulated variationally as:
find $u \in H_0^1(\Omega)$, such that

$$a(u,v) = \int_\Omega (\varepsilon\, \rho(\mathbf{x}) \nabla u \cdot \nabla v + c\, u\, v)\, d\mathbf{x} = f(v), \quad v \in H_0^1(\Omega), \tag{3.1}$$

where c, ε are non-negative real coefficients. $H^1(\Omega)$ is the space of square summable functions with square summable first derivatives, and $H_0^1(\Omega)$ its subspace of functions that vanish on $\partial\Omega$. The functional $f(\cdot)$ belongs to the dual space $H^{-1}(\Omega)$. Here $\mathbf{x} = (x, y, z)$ denotes the position vector. The coefficient $\rho(\mathbf{x}) > 0$ can be discontinuous, with very different values for different subregions of Ω, but we allow it to vary only moderately within each subregion. Without decreasing the generality of our results, we will only consider the piecewise constant case i.e. $\rho(\mathbf{x}) = \rho_i, \mathbf{x} \in \Omega_i$. Later we consider the purely diffusion problem derived from (3.1) as a model problem to derive the condition number bound for the balancing Neumann-Neumann and one-level FETI preconditioners.

3.2.2. *hp* finite element approximations

We now specify a particular choice of finite element spaces. Let $\mathcal{T}$ be a mesh consisting of affinely mapped cubes. Given a polynomial degree $k \geq 1$, we consider the following finite element spaces

$$X = X^k(\Omega; \mathcal{T}) = \left\{ u \in H_0^1(\Omega) \,|\quad u_{|_K} \in \mathbb{Q}_k(K),\ K \in \mathcal{T} \right\}. \tag{3.2}$$

Here $\mathbb{Q}_k(K)$ is the space of polynomials of maximum degree k in each variable on K. In the following, we may drop the reference to k, Ω, and/or $\mathcal{T}$ whenever there is no confusion.

In this analysis, interpolating Lagrange polynomials on Gauss-Lobatto nodes are used as a particular nodal basis of $X^k(\Omega;\mathcal{T})$. The set of Gauss-Lobatto points $GLL(k)$ is the set of (distinct and real) zeros of $(1-x^2)L_k'(x)$, with L_k the Legendre polynomial of degree k (cf. [35, Sect. 3]) and the quadrature formula based on $GLL(k)$ has order $2k-1$. In this work, quadrature formulas based on $GLL(k)$ are chosen. Given the nodes $GLL(k)^3$ on the reference element $\hat{Q} = (-1,1)^3$, our basis functions on $\mathbb{Q}_k(\hat{Q})$ are defined as tensor products of k-th order Lagrange interpolating polynomials on $GLL(k)$. More details on spectral element methods can be found in, e.g., [35].

We always assume that the meshes are *regular*, i.e., the intersection between neighboring elements is either a vertex, or an edge, or a face that is common to both elements. A finite element approximation of (3.1) consists of finding $u \in X$, such that

$$a(u,v) = f(v), \quad v \in X. \tag{3.3}$$

3.2.3. Geometrically refined meshes

We now introduce a class of geometrically graded meshes. They are determined by a mesh grading factor $\sigma \in (0,1)$ and a refinement level $n \geq 0$. The number of layers is $n+1$ and the thinnest layer has a width proportional to σ^n (see Figures 3.1 and 3.2 for illustrations in three and two dimensions, respectively). Robust exponential convergence of *hp* finite element approximations is achieved if n is suitably chosen. For singularity resolution, n is required to be proportional to the polynomial degree k; see [12, 17]. For boundary layers, the width of the thinnest layer mesh needs to be comparable to that of the boundary layer; see [178, 220, 221].

A geometric boundary layer mesh $\mathcal{T} = \mathcal{T}_{bl}^{n,\sigma}$ is obtained as tensor products of meshes that are geometrically refined towards the faces. The mesh $\mathcal{T}_{bl}^{n,\sigma}$ is built from an initial shape-regular macro-triangulation $\mathcal{T}^0$, possibly consisting of just one element, which is successively refined. Every macroelement can be refined isotropically or anisotropically as a face, edge, or corner patch. A refinement towards a corner is shown in left part of Figure 3.1. We refer the reader to [242, 243] for more details on the construction of these meshes. Note that the mesh aspect ratio is equal to $\sigma^{-n} \sim \sigma^{-k}$, since n needs to be comparable to k for exponential convergence. A geometric boundary layer mesh $\mathcal{T}$ satisfies the following two properties

Property 3.1. $\mathcal{T}$ is obtained from an initial shape-regular coarse mesh $\mathcal{T}^0$ (called macromesh) by local isotropic or anisotropic refinement.

Property 3.2. Anisotropic refinement is always performed towards the boundary $\partial\Omega$ of the computational domain Ω and never towards the interior.

Figures 3.1 (left part) and 3.2 highlight these features.

3.2.4. Domain partitioning and assembly phase

Iterative substructuring methods rely on a non-overlapping partition of Ω, $\mathcal{T}^{DD} = \{\Omega_i\}$, into substructures. Let M denote the number of substructures with H_i the diameter of Ω_i and $H = \max(H_i)$ the maximum of their diameters. A subdomain Ω_i is called *floating* if the intersection of $\partial\Omega_i$ with $\partial\Omega$ is empty. We recall that we have only considered

the case of Dirichlet boundary conditions. We define the boundaries $\Gamma_i = \partial\Omega_i \setminus \partial\Omega$ and the interface Γ as their union. The sets of Gauss-Lobatto nodes and the corresponding degrees of freedom on $\partial\Omega_i$, Γ_i, Γ, and $\partial\Omega$ are denoted by $\partial\Omega_{i,h}$, $\Gamma_{i,h}$, Γ_h, and $\partial\Omega_h$, respectively.

The main geometric assumption on the substructures is that they are *shape-regular*. This property appears to be essential to obtain the condition number bound. Indeed, Property 3.1 allows us to satisfy this condition easily by choosing the macromesh as the subdomain partition

$$\mathcal{T}^{DD} = \mathcal{T}^0.$$

A consequence of Property 3.2 is then that, when two substructures share an interior vertex, the local meshes are shape-regular in the neighbourhood of this vertex, since anisotropic refinement is only performed towards the boundary $\partial\Omega$.

3.3. Preconditioner in the primal space: the balancing Neumann-Neumann method

In this section, we describe and analyse a hybrid Schwarz algorithm known as the balancing Neumann-Neumann method [172]. We first derive the method in a matrix form, give the condition number bound and finally provide numerical experiments supporting the theory.

3.3.1. Derivation

After subassembling, the stiffness matrix A is reordered according to the domain decomposition partitioning. The nodal points interior to the substructures (subset I) are ordered first, followed by those on the interface Γ (subset Γ). Similarly, for the local stiffness matrix relative to a substructure Ω_i, we have

$$A^{(i)} = \begin{pmatrix} A_{II}^{(i)} & A_{I\Gamma}^{(i)} \\ A_{\Gamma I}^{(i)} & A_{\Gamma\Gamma}^{(i)} \end{pmatrix}.$$

First, the unknowns in the interior of the substructures are eliminated by block Gaussian elimination. Unknowns on $\partial\Omega_i \cap \partial\Omega$ are treated as interior and they are also eliminated. In this step, the Schur complement $S = S_{NN}$ with respect to the interior variables is formed. The resulting linear system for the nodal values on Γ can be written as

$$S_{NN}\, u_\Gamma = g_\Gamma. \tag{3.4}$$

Given the local Schur complement associated with the substructure Ω_i and the local right-hand side

$$S_i = A_{\Gamma\Gamma}^{(i)} - A_{\Gamma I}^{(i)} A_{II}^{(i)^{-1}} A_{I\Gamma}^{(i)} \qquad g_{\Gamma_i} = b_{\Gamma_i} - A_{\Gamma I}^{(i)} A_{II}^{(i)^{-1}} b_I^{(i)}, \tag{3.5}$$

the global Schur complement and the corresponding right-hand side g_Γ can be written as

$$S = S_{NN} = \sum_{i=1}^{M} R_i^T S_i R_i \qquad g_\Gamma = \sum_{i=1}^{M} R_i^T g_{\Gamma_i}\,, \tag{3.6}$$

where the restriction matrix R_i is a matrix of zeros and ones which extracts the variables on the local interface Γ_i from a vector of nodal values on Γ. The balancing Neumann-Neumann preconditioner $\hat{S}^{-1}$ [172] provides a preconditioned operator P_{NN} of the following form

$$P_{NN} = \hat{S}^{-1} S_{NN} = P_0 + (I - P_0)(\sum_{i=1}^{M} P_i)(I - P_0). \tag{3.7}$$

Here P_0 is associated with a low dimensional global coarse problem, whereas each operator P_i is associated with one substructure. More precisely, the local operators P_i are defined as

$$P_i = R_i^T D_i S_i^\dagger D_i R_i S_{NN}, \tag{3.8}$$

where the matrices D_i are diagonal and $S_i^\dagger$ denotes either the inverse of S_i, if S_i is non-singular as for subdomains that touch $\partial\Omega$, or a pseudoinverse of S_i, if S_i is singular as for floating domains. In order to define the matrices $\{D_i\}$, we need to introduce a scaling function $\delta_i^\dagger$, which is a finite element function defined on the boundary $\partial\Omega_i$; cf. [81, 82, 172, 198, 219]. To do so, it is enough to assign its values at the nodes in $\Gamma_{i,h}$. It is defined for $\gamma \in [1/2, \infty)$ and, is determined by a sum of contributions from Ω_i and its relevant nearest neighbours,

$$\delta_i^\dagger(x_l) = \frac{\left(a_{ll}^{(i)}\right)^\gamma}{\sum_{j\in\mathcal{N}_{x_l}} \left(a_{ll}^{(j)}\right)^\gamma}, \quad x_l \in \Gamma_{i,h}\,, \tag{3.9}$$

where $a_{ll}^{(i)}$ denotes the l-th element of the diagonal of the local stiffness matrix $A^{(i)}$ and $\mathcal{N}_{x_l}$, $x_l \in \Gamma_h$, is the set of indices j of the subregions such that $x_l \in \Gamma_{j,h}$. We have chosen $\gamma = 1$ for the numerical experiments of Section 3.3.4. Let D_i be the diagonal matrix with elements $\delta_i^\dagger(x)$ corresponding to the nodes in $\Gamma_{i,h}$. The coarse space is defined as

$$V_0 = \text{span}\{R_i^T \delta_i^\dagger\},$$

where the span is taken over at least the floating subdomains. We denote by R_0^T the prolongation from the coarse to the global space. In analogy with (3.8), the coarse operator P_0 is defined as

$$P_0 = R_0^T S_0^{-1} R_0 S_{NN}, \tag{3.10}$$

where $S_0 = R_0 S_{NN} R_0^T$ denotes the restriction of S_{NN} to that coarse space. We refer the reader to [243] for more details.

3.3.2. Condition number bound

We recall here the two results related to the condition number bounds in two and three dimensions, respectively.

Two-dimensional setting

In the case of purely diffusive problems corresponding to $c = 0$ in (3.1) for a similar problem in two dimensions ([242]), a bound for the condition number of the preconditioned operator P_{NN} restricted to the subspace $\mathcal{R}(I - P_0)$, to which the iterates are confined, has been proved in [242] for the case of exact solvers for Neumann and Dirichlet problems. We have

$$\kappa(P_{NN}) \leq C\,(1-\sigma)^{-4}\left(1+\log\left(\frac{k}{1-\sigma}\right)\right)^2, \tag{3.11}$$

where the constant C is independent of the spectral polynomial degree k, the level of refinement n, the mesh grading factor σ, the coefficients ε and ρ, and the diameters of the substructures H_i.

Three-dimensional setting

Exact variant In the case of purely diffusive problems corresponding to $c = 0$ in (3.1), a bound for the condition number of the preconditioned operator P_{NN} restricted to the subspace $\mathcal{R}(I - P_0)$, to which the iterates are confined, has been proved in [243] for the case of exact solvers for Neumann and Dirichlet problems. We have

$$\kappa(P_{NN}) \leq C\,(1-\sigma)^{-6}\left(1+\log\left(\frac{k}{1-\sigma}\right)\right)^2, \tag{3.12}$$

where the constant C is independent of the spectral polynomial degree k, the level of refinement n, the mesh grading factor σ, the coefficients ε and ρ, and the diameters of the substructures H_i. In relations (3.11) and (3.12), we note that $\kappa(P_{NN})$ does not depend on the number of substructures or the aspect ratio of the mesh and only depends polylogarithmically on the spectral polynomial degree k as in the p version on shape-regular meshes [198].

Inexact variant In the case of large local problems that often arise in three dimensions, exact solvers may be too expensive in terms of computational operations and/or memory requirements. Approximate local solvers for both Neumann and Dirichlet problems must then be used. This new setting often leads to a *variable* domain decomposition preconditioner. Hence, *flexible* Krylov subspace methods as an outer method are required as discussed later in Chapter 4. We note that the bound (3.12) is no longer valid since approximate solvers are used. Hence, we will investigate numerically the behaviour of the condition number of the preconditioned operator in such a situation in Section 3.3.4. We refer the reader to [226, Section 4.4] and [246, Section 4.3] for further references related to the analysis of inexact variants of domain decomposition preconditioners.

3.3.3. Algorithm

According to (3.4) and (3.7), the preconditioned system can be written in the following form

$$P_{NN}u = \hat{S}^{-1}g_\Gamma. \tag{3.13}$$

Since P_0 is a projection, we have

$$P_0(I - P_0) = 0.$$

Thus a decomposition of the exact solution u of (3.13) as

$$u = P_0u + w, \qquad P_0u = R_0^T S_0^{-1} R_0 g_\Gamma, \tag{3.14}$$

with $w \in \mathcal{R}(I - P_0)$, leads to the following new formulation of (3.13)

$$(I - P_0)(\sum_{i=1}^{M} P_i)(I - P_0)w = \hat{S}^{-1}g_\Gamma - P_0u, \qquad w \in \mathcal{R}(I - P_0). \tag{3.15}$$

One can easily check that $S_{NN}P_0 = P_0^T S_{NN}$, and thus the matrix in (3.15) can also be written as

$$(I - P_0)(\sum_{i=1}^{M} P_i)(I - P_0) = \left[(I - P_0)(\sum_{i=1}^{M} R_i^T D_i S_i^\dagger D_i R_i)(I - P_0^T)\right] S_{NN},$$

which gives the expression of the preconditioner $\hat{S}^{-1}$. Consequently, the balancing Neumann-Neumann method reduces to a projected preconditioned conjugate gradient method in the space $\mathcal{R}(I - P_0)$ applied to the system

$$\left[(I - P_0)(\sum_{i=1}^{M} R_i^T D_i S_i^\dagger D_i R_i)(I - P_0^T)\right] S_{NN}\, w = \hat{S}^{-1}g_\Gamma - P_0u \tag{3.16}$$

if an initial guess $u_0 = P_0u + \tilde{w}$, with $\tilde{w} \in \mathcal{R}(I - P_0)$, is chosen. The projected conjugate gradient method is presented in Algorithm 3.1. In this algorithm, $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. Thanks to (3.14) and to the choice of u_0, the first projection step, corresponding to the application of $I - P_0^T$, can be omitted in practice.

We remark that the matrices S_{NN} and $S_i^\dagger$ do not need to be calculated in practice. The action of S_{NN} on a vector requires the solution of a Dirichlet problem on each substructure (application of the inverse of $A_{II}^{(i)}$), while the action of $S_i^\dagger$ can be calculated by applying a pseudo-inverse of $A^{(i)}$ to a suitable vector, corresponding to the solution of a Neumann problem; see [226, Chap 4.]. Thus one step of the algorithm in Algorithm 3.1 involves one application of P_0, the solution of local Neumann problems on each substructure ($S_i^\dagger$) and the solution of local Dirichlet problems (S_{NN}). Since the application of P_0 also involves an application of S_{NN} and the solution of a coarse problem, the total amount of work per step is given by one Neumann and two Dirichlet problems on each substructure and one coarse problem.

Algorithm 3.1 Implementation of the balancing Neumann-Neumann method as a projected preconditioned conjugate gradient method.

Input: Assume that the following is given:

- M ▷ number of subdomains
- S_{NN} ▷ global Schur complement operator (3.6)
- S_i for $i \in \{1, \ldots, M\}$ ▷ local Schur complement operator (3.5)
- P_i for $i \in \{1, \ldots, M\}$ ▷ local projection operator (3.8)
- R_i for $i \in \{1, \ldots, M\}$ ▷ local restriction operator
- P_0 ▷ coarse level projection operator (3.10)
- R_0 ▷ coarse restriction operator
- $S_0 = R_0 S_{NN} R_0^T$ ▷ coarse level Schur complement operator

$u_0 = R_0^T S_0^{-1} R_0 g_\Gamma + \tilde{w}, \quad \tilde{w} \in \mathcal{R}(I - P_0)$
$q_0 = g_\Gamma - S_{NN}\, u_0$
for $j = 1, \ell$ **do**
  $w_{j-1} = (I - P_0^T) q_{j-1}$ ▷ projection
  $z_{j-1} = \sum_{i=1}^{M} R_i^T D_i S_i^\dagger D_i R_i\, w_{j-1}$ ▷ preconditioning
  $y_{j-1} = (I - P_0) z_{j-1}$ ▷ projection
  $\beta_j = \langle y_{j-1}, w_{j-1} \rangle / \langle y_{j-2}, w_{j-2} \rangle \quad [\beta_1 = 0]$
  $p_j = y_{j-1} + \beta_j p_{j-1} \quad [p_1 = y_0]$
  $\alpha_j = \langle y_{j-1}, w_{j-1} \rangle / \langle p_j, S_{NN}\, p_j \rangle$
  $u_j = u_{j-1} + \alpha_j p_j$
  $q_j = q_{j-1} - \alpha_j S_{NN}\, p_j$
end for

3.3.4. Numerical results

The purpose of this section is to provide a single numerical experiment, in order to validate the theoretical analysis related to the condition number bound on selected medium-size problems. This numerical experiment targets the efficiency of the balancing Neumann-Neumann preconditioner for a Laplace problem defined on a boundary layer mesh (exhibiting a corner refinement). The conjugate gradient iteration is stopped after a reduction of the Euclidean norm of the initial residual of 10^{-14} and homogeneous boundary conditions have been used. An extensive numerical study is presented elsewhere; see [241].

Laplace problem on a boundary layer mesh

We consider approximations on the unit cube $\Omega = (0,1)^3$. We choose $\rho \equiv 1$ and the right-hand side $f \equiv 1$. The macromesh $\mathcal{T}_m$ consists of $N \times N \times N$ cubic substructures. Geometric refinement is performed towards the three edges $x = 0$, $y = 0$, and $z = 0$, with $\sigma = 0.5$; see Figure 3.1, left. Given a polynomial degree k, we choose $n = k$ as is required for robust exponential convergence; see, e.g., [12, 17].

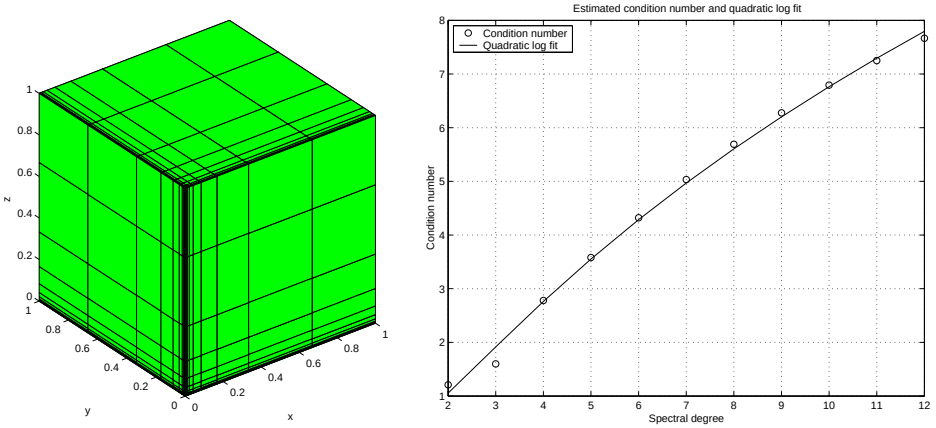


Figure 3.1.: Geometric refinement towards one corner ($N = 3$, $\sigma = 0.5$, and $n = 6$), left, and estimated condition numbers (circles) from Table 3.1 (inexact variant) and least-square second order logarithmic polynomial fit (solid line) versus k, right. Figures 3 (left) and 1 (right) of [241].

We note that even for moderate values of k and N, large linear systems are obtained; see Tables 3.1 and 3.2. Huge local blocks need to be inverted, both for the application of S (solution of local Dirichlet problems) and the preconditioner (solution of local Neumann problems). Thus we have employed approximate solvers for local Dirichlet and Neumann problems. We refer to [226, Sect. 4.4] for details on the implementation. In particular, we have used a conjugate gradient iteration with an incomplete Cholesky factorization with drop tolerance 10^{-3} for all local problems. The iteration is stopped after a reduction of the initial residual of a factor 10^{-3} or after 20 iteration steps. In the sequel, we denote by NN (inexact) the resulting balancing Neumann-Neumann method with this strategy for the approximate solvers. An exact variant denoted by NN (exact) is derived, when solving all the local subproblems now up to machine precision with the same iterative solver as in the inexact case. Our numerical results show that the theoretical bounds for the case of exact solvers in Lemma 3.4 remain valid in this case; see Tables 3.1 and 3.2.

For a fixed partition into substructures with $N = 3$, Table 3.1 shows the size of the original problem (n), the iteration count (It), the estimated maximum and minimum eigenvalues (λ_{max} and λ_{min} obtained by calculating the eigenvalues of the Hessenberg tridiagonal matrix constructed during the conjugate gradient iteration), and the condition number for different values of k for both inexact and exact variants. We note that the minimum eigenvalue is close to one when using inexact solvers; see Lemma 3.4. In addition, a moderate growth of the maximum eigenvalue is observed with k; such growth is consistent with the quadratic bound in Lemma 3.6; see Figure 3.1, right. Using inexact solvers for the local subproblems causes a moderate increase in terms of number of iterations. Nevertheless, quite satisfactory condition numbers are still obtained, see Table 3.1.

Fixed number of substructures ($N \times N \times N = 3 \times 3 \times 3$)									
		NN (inexact)				NN (exact)			
k	n	It	λ_{max}	λ_{min}	$\kappa(P_{NN})$	It	λ_{max}	λ_{min}	$\kappa(P_{NN})$
2	1331	15	1.8379	1	1.8379	13	1.6255	1.00002	1.6255
3	6859	20	2.8165	0.99997	2.8166	18	2.8165	1.00001	2.8161
4	24389	25	3.9507	0.99947	3.9528	21	3.9506	1.00002	3.9498
5	68921	29	5.1507	0.99799	5.1611	25	5.1507	1.00002	5.1493
6	166375	34	6.3675	0.99801	6.3803	28	6.3675	1.00002	6.3658
7	357911	38	7.5082	0.99395	7.5540	32	7.5067	1.00002	7.5065
8	704969	40	8.5298	0.99574	8.5663	34	8.5064	1.00002	8.5062

Table 3.1.: Conjugate gradient method for the global system with Neumann-Neumann preconditioner with inexact and exact solvers: iteration counts, maximum and minimum eigenvalues, and condition numbers, versus the polynomial degree, for the case of a fixed partition. The size of the original problem is also reported. Table 2 of [243].

We next consider the same problem and fix the polynomial degree as $k = 4$. Table 3.2 shows the results for different values of N. In both variants, the iteration counts and the smallest and largest eigenvalues appear to be bounded independently of the number of subdomains. We note that, when the number of subdomains becomes large, the number of iterations to reach the convergence criterion for both variants is nearly identical, leading to a *scalable* method.

3.4. Preconditioners in the dual space: the one-level FETI method

One-level FETI methods were first introduced in [103]. Since then, considerable research has been done on FETI methods and many variants and improvements have been proposed. We refer to [104] for a detailed introduction and to [158, 175] for the analysis of one-level FETI methods.

3.4.1. Derivation

For the sake of brevity, we only present the one-level FETI method in the case of purely diffusive problems as in Section 3.3. We refer to [240] for details in the case of reaction-diffusion problems, i.e., when the local matrices $A^{(i)}$ are invertible. Instead of solving the Schur complement system (3.4), a FETI method uses a space of discontinuous functions across the interface Γ. The continuity of the solution is then enforced by using a

Fixed spectral degree $k = 4$									
		NN (inexact)				NN (exact)			
N	n	It	λ_{max}	λ_{min}	$\kappa(P_{NN})$	It	λ_{max}	λ_{min}	$\kappa(P_{NN})$
2	15625	18	2.6417	0.99929	2.6436	15	2.6412	1.0003	2.6406
3	24389	25	3.9507	0.99947	3.9528	21	3.9506	1.0002	3.9498
4	35937	28	4.1084	0.99934	4.1111	25	4.1082	1.0002	4.1074
5	50653	29	4.1378	0.99940	4.1402	26	4.1375	1.0002	4.1369
6	68921	30	4.1492	0.99945	4.1515	28	3.5746	1.0002	3.5741
7	91125	30	4.1555	0.99952	4.1575	28	3.6133	1.0001	3.6128
8	117649	30	4.1593	0.99955	4.1612	29	3.6289	1.0001	3.6284
9	148877	30	4.1618	0.99962	4.1634	29	3.6475	1.0001	3.6470
10	185193	30	4.1636	0.99970	4.1648	29	3.6582	1.0001	3.6577

Table 3.2.: Conjugate gradient method for the global system with Neumann-Neumann preconditioner with inexact and exact solvers: iteration counts, maximum and minimum eigenvalues, and condition numbers, versus the number of substructures, using a fixed polynomial degree and partitions into $N \times N \times N$ substructures. The size of the original problem is also reported. Table 3 of [243].

vector of Lagrange multipliers and this leads to the saddle point formulation

$$\left.\begin{aligned} S_F u_F + B^T \lambda &= g_F \\ B u_F &= 0 \end{aligned}\right\}, \tag{3.17}$$

with

$$u_F = \begin{bmatrix} u^{(1)} \\ u^{(2)} \\ \vdots \\ u^{(M)} \end{bmatrix}, \; S_F = \begin{bmatrix} S_1 & O & \cdots & O \\ O & S_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & O \\ O & \cdots & O & S_M \end{bmatrix}, \; g_F = \begin{bmatrix} g^{(1)} \\ g^{(2)} \\ \vdots \\ g^{(M)} \end{bmatrix}$$

where each diagonal block of S_F is a Schur complement matrix of the form (3.5) and B a matrix consisting of $(-1, 0, 1)$ that enforces the continuity of the solution at the interfaces between the substructures. Here only non-redundant Lagrange multipliers have been considered; thus the matrix B has full rank. We refer to [158] for the analysis in the case of redundant Lagrange multipliers. In addition, we denote by R the full column rank matrix built from all the non-void null space elements of S_F, i.e.,

those S_i corresponding to floating subdomains

$$R = \begin{bmatrix} r_1 & O & \cdots & O \\ O & r_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & O \\ O & \cdots & O & r_{M_f} \end{bmatrix}$$

where M_f denotes the number of floating domains. In fact, for the purely diffusive model problem (3.1) the columns of R span the kernel of S_F denoted by $\mathcal{N}(S_F)$. We also define $G = BR$. In the next step, we eliminate the primal variable u_F in (3.17) and derive an equation for the Lagrange multiplier λ only. This leads to

$$u_F = {S_F}^{\dagger}(g_F - B^T\lambda) + R\alpha, \quad (g_F - B^T\lambda) \perp \mathcal{N}(S_F),$$

with $S_F^{\dagger}$ denoting a pseudo-inverse of S_F and

$$\left.\begin{aligned} F\lambda - G\alpha &= d \\ G^T\lambda &= e \end{aligned}\right\}, \tag{3.18}$$

with $F = B{S_F}^{\dagger}B^T$, $d = B{S_F}^{\dagger}g_F$, and $e = R^T g_F$. After introducing a suitable orthogonal projection operator P onto the orthogonal complement of $\mathcal{R}(G)$ and a preconditioner M_F^{-1} (both defined below), the one-level FETI method reduces to the preconditioned conjugate gradient method applied in the space of Lagrange multipliers to the following system

$$PM_F^{-1}P^TF\lambda = PM_F^{-1}P^Td, \tag{3.19}$$

with an initial approximation λ_0 that satisfies the second equation of (3.18). We can choose

$$\lambda_0 = QG(G^TQG)^{-1}R^Tg_F + \tilde{w}, \qquad \tilde{w} \in \mathcal{R}(P)\,, \tag{3.20}$$

where Q is a symmetric invertible matrix to be chosen. Here P is an orthogonal projection operator defined as

$$P = I - QG(G^TQG)^{-1}G^T. \tag{3.21}$$

Many choices have been proposed for the preconditioner M_F^{-1} and the matrix Q. The choice

$$M_F^{-1} = (BD^{-1}B^T)^{-1}BD^{-1}S_FD^{-1}B^T(BD^{-1}B^T)^{-1}, \quad Q = M_F^{-1} \tag{3.22}$$

ensures a condition number that is independent of the jumps in the coefficients; see [158]. Here D is a block diagonal matrix: each block D_i corresponds to one substructure Ω_i and is equal to the local scaling matrix introduced in Section 3.3.1.

3.4.2. Condition number bound

For the sake of brevity, we prefer to give without proof the condition numbers related to the one-level FETI preconditioner. We refer the reader to [242] for a complete proof of the condition number bound in two dimensions. In three dimensions, the condition number bound directly follows from the analysis of the balancing Neumann-Neumann preconditioner [243] (partly given in Section 3.5) and the relations between the balancing Neumann-Neumann and one-level FETI methods established in [158].

Two-dimensional setting

We denote by $P_F = PM_F^{-1}P^TF$ the preconditioned operator in system (3.19). A bound for the condition number of P_F restricted to the appropriate subspace $\mathcal{R}(P)$ to which the iterates are confined has been proved in [242] for the case $(\rho_x, \rho_y) = (\rho, \rho)$, $(\varepsilon_x, \varepsilon_y) = (1, 1)$ and $c = 0$ in two dimensions

$$\kappa(P_F) \leq C\,(1-\sigma)^{-4}\left(1+\log\left(\frac{k}{1-\sigma}\right)\right)^2. \tag{3.23}$$

Three-dimensional setting

We denote by $P_F = PM_F^{-1}P^TF$ the preconditioned operator in system (3.19). A bound for the condition number of P_F restricted to the appropriate subspace $\mathcal{R}(P)$ to which the iterates are confined reads as for the case in the purely diffusive case ($c = 0$) in three dimensions

$$\kappa(P_F) \leq C\,(1-\sigma)^{-6}\left(1+\log\left(\frac{k}{1-\sigma}\right)\right)^2. \tag{3.24}$$

In relations (3.23) and (3.24), we stress the fact that the constant in the estimate is independent of k, n, σ, the coefficients ρ and the diameters H_i of the substructures. Note that $\kappa(P_F)$ does not depend on the number of substructures or the aspect ratio of the mesh and only depends polylogarithmically on the spectral polynomial degree.

3.4.3. Algorithm

The one-level FETI method is a projected preconditioned conjugate gradient method in the space of Lagrange multipliers $\mathcal{R}(P)$ applied to the system (3.19) with an initial approximation chosen as in (3.20). This algorithm is given in Algorithm 3.2. We note that, as opposed to the balancing Neumann-Neumann algorithm, the first projection step (application of P^T) cannot be omitted in practice.

We remark that F and M_F^{-1} do not need to be calculated in practice. The action of M_F^{-1} on a vector basically requires the solution of a Dirichlet problem on each substructure (application of S_F, and thus the S_i). Indeed, the matrix $BD^{-1}B^T$ is block diagonal: each block corresponds to a node on Γ and its dimension is equal to the number of constraints imposed on that node by the second of (3.17): it can then be easily inverted. The action of F can be calculated by solving Neumann problems on the substructures (application of the pseudoinverses $S_i^\dagger$). Finally, applications of P and P^T

Algorithm 3.2 Implementation of the one-level FETI method as a projected preconditioned conjugate gradient method.

Input: Assume that the following is given:

- M ▷ number of subdomains
- F ▷ $BS_F{}^{\dagger}B^T$ operator
- P ▷ orthogonal projection operator (3.21)
- M_F^{-1} ▷ FETI preconditioner (3.22)
- Q ▷ symmetric operator (3.22)

1: $\lambda_0 = QG(G^TQG)^{-1}R^Tg_F + \tilde{w}, \quad \tilde{w} \in \mathcal{R}(P)$
2: $q_0 = d - F\lambda_0$
3: **for** $j = 1, \ell$ **do**
4: $\quad w_{j-1} = P^Tq_{j-1}$ ▷ projection
5: $\quad z_{j-1} = M_F^{-1}w_{j-1}$ ▷ preconditioning
6: $\quad y_{j-1} = Pz_{j-1}$ ▷ projection
7: $\quad \beta_j = \langle y_{j-1}, w_{j-1}\rangle / \langle y_{j-2}, w_{j-2}\rangle \quad [\beta_1 = 0]$
8: $\quad p_j = y_{j-1} + \beta_j p_{j-1} \quad [p_1 = y_0]$
9: $\quad \alpha_j = \langle y_{j-1}, w_{j-1}\rangle / \langle p_j, Fp_j\rangle$
10: $\quad \lambda_j = \lambda_{j-1} + \alpha_j p_j$
11: $\quad q_j = q_{j-1} - \alpha_j Fp_j$
12: **end for**

are required at each step and involve the solution of two coarse problems (application of $(G^TQG)^{-1}$) and two additional applications of M_F^{-1}. The total amount of work per step requires the solution of one Neumann and three Dirichlet problems on each substructure and two coarse problems. We note however that the choice $Q = M_F^{-1}$ appears to be necessary only if the coefficients have large jumps; see [158].

3.4.4. Numerical results

The purpose of this section is to provide a single numerical experiment in order to validate the theoretical analysis related to the condition number bound. The numerical experiment targets the efficiency of the Neumann-Neumann and one-level FETI preconditioners for a Laplace problem defined on a boundary layer mesh (exhibiting a corner refinement) in two dimensions. The conjugate gradient iteration is stopped after a reduction of the Euclidean norm of the initial residual of 10^{-14} (this rather strict stopping criterion allows a possible comparison between the proposed balancing Neumann-Neumann and one-level FETI preconditioner, since the primal solution in the one-level FETI formulation is only continuous at convergence) and homogeneous boundary conditions have been used. An extensive numerical study is presented in [240].

Laplace problem on a boundary layer mesh

As in Section 3.3.4, we fix a macromesh $\mathcal{T}^{DD} = N_x \times N_y = 3 \times 3$ and investigate the dependence of the condition number on the spectral polynomial degree. The geometrically refined grid $\mathcal{T}$ contains $(3+k) \times (3+k)$ elements; see Figure 3.2 for the case $k = 6$.

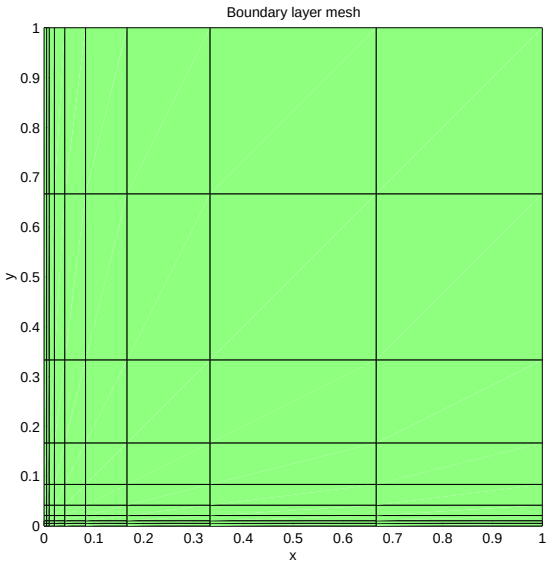


Figure 3.2.: Geometric refinement towards one corner ($N = 3$, $\sigma = 0.5$, and $n = 6$). Figure 6 of [240].

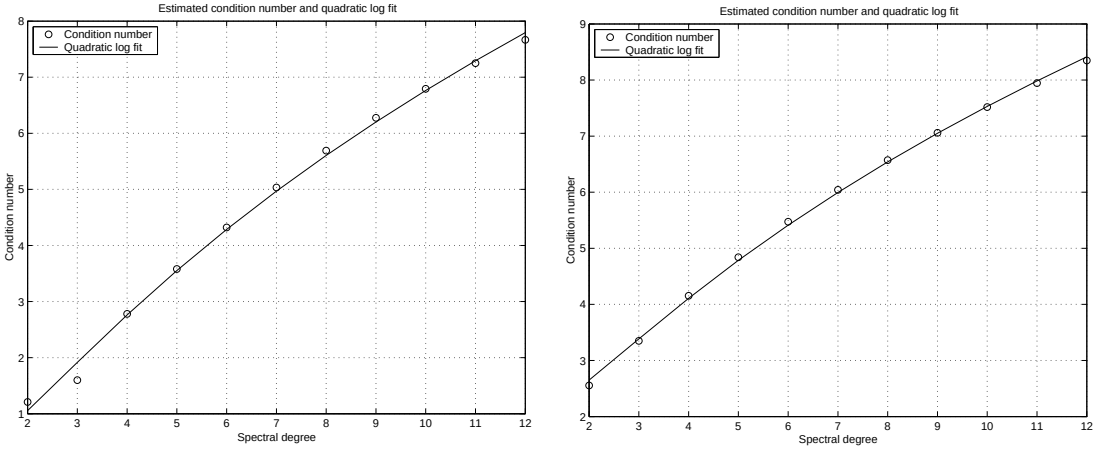
Tables 3.3 and 3.4 collect the results for the balancing Neumann-Neumann and one-level FETI preconditioners, respectively. Note the large condition numbers of unpreconditioned operators $\kappa(S_{NN})$ and $\kappa(F)$ for large k. We expect $\kappa(S_{NN})$ to grow as $k\,\sigma^{-(\beta\,k)}$, for a fixed partition. As expected (see Figure 3.3), $\kappa(P_{NN})$ and $\kappa(P_F)$ only grow as the square of the logarithm of the spectral degree. This is in agreement with the bounds in (3.11) and (3.23). These results show that the condition numbers are independent of the aspect ratio of the mesh for this problem. We stress the fact that the original Schur complement has a condition number that grows *exponentially* with k, while our preconditioners provide a condition number that only grows *logarithmically* with k.

3.5. Basis of a theory

We provide in this section the basis of a theory required to derive the condition number bound related to the balancing Neumann-Neumann preconditioner for *hp* finite element approximations on anisotropic meshes in three dimensions. We refer the reader to [226, Chapter 5], [246, Chapter 2] for a detailed introduction to the abstract theory of Schwarz

Fixed number of substructures $N_x \times N_y = 3 \times 3$								
	No preconditioning				NN			
k	It	λ_{max}	λ_{min}	$\kappa(S_{NN})$	It	λ_{max}	λ_{min}	$\kappa(P_{NN})$
2	18	13.09	0.47009	27.8466	8	1.2093	1	1.2093
3	39	23.584	0.27906	84.5135	10	1.5992	1	1.5991
4	71	43.421	0.19866	218.5623	12	2.7807	1	2.7806
5	118	82.489	0.15446	534.0585	13	3.5809	1.0001	3.5806
6	185	160.4	0.12649	1268.0820	14	4.321	1.0001	4.3204
7	272	315.84	0.10716	2947.3406	15	5.034	1.0002	5.0331
8	344	625.76	0.092981	6729.9791	17	5.6913	1.0001	5.6906
9	424	1243.8	0.082121	15145.9124	17	6.2769	1.0001	6.2759
10	512	2476.8	0.073532	33683.7624	17	6.7937	1.0002	6.7924
11	608	4937.9	0.066568	74178.6450	18	7.2527	1.0002	7.2510
12	712	9852.1	0.060824	161978.5169	19	7.6679	1.0002	7.6660

Table 3.3.: Conjugate gradient method for the global system with balancing Neumann-Neumann preconditioner: iteration counts, maximum and minimum eigenvalues, and condition numbers, versus the polynomial degree, using a fixed number of substructures and partition into 3×3 substructures. Table 10 of [240].



balancing Neumann-Neumann one-level FETI

Figure 3.3.: Laplace problem on a boundary layer mesh. Fixed partition 3×3. Estimated condition numbers (circles) and least-square second order logarithmic polynomial (solid line) versus the spectral degree for the balancing Neumann-Neumann method (left) and the one-level FETI method (right). Figure 7 of [240].

Fixed number of substructures $N_x \times N_y = 3 \times 3$								
	No preconditioning				One-level FETI			
k	It	λ_{max}	λ_{min}	$\kappa(F)$	It	λ_{max}	λ_{min}	$\kappa(P_F)$
2	40	11.092	0.3057	36.2830	13	2.5551	1.0002	2.5545
3	69	16.1727	0.1696	95.3514	14	3.3502	1.0003	3.3490
4	101	21.5275	0.09212	233.6839	16	4.15468	1.00025	4.1536
5	157	27.0024	0.04849	556.8545	17	4.8423	1.0005	4.8399
6	214	32.4910	0.0249	1302.8742	18	5.4769	1.0006	5.4732
7	280	37.9728	0.01268	2994.9336	18	6.0492	1.0013	6.4125
8	352	43.4460	0.00649	6688.1144	19	6.5801	1.0012	6.5721
9	432	48.9158	0.0048	10140.5428	20	7.0699	1.0014	7.0597
10	520	54.38492	0.002428	22398.1226	20	7.5287	1.0013	7.5183
11	616	59.8554	0.0012	48165.1532	21	7.9582	1.0016	7.9449
12	720	65.3276	0.00065	99925.7334	20	8.3638	1.0018	8.3484

Table 3.4.: Conjugate gradient method for the global system with one-level FETI preconditioner: iteration counts, maximum and minimum eigenvalues, and condition numbers, versus the polynomial degree, using a fixed number of substructures and partition into 3×3 substructures. Table 11 of [240].

methods which relies on variational forms. The interested reader can find the complete derivation in [243] (see also Appendix B.2).

3.5.1. Local meshes, local bilinear forms and local extension operators

First, we introduce local meshes and local bilinear forms to be used later.

Local meshes When restricted to the subdomain Ω_i, the global triangulation $\mathcal{T}$ determines a local mesh $\mathcal{T}_i$. In three dimensions this mesh can be of four types: face, edge, corner, or consisting of just one element. We define the local spaces $X_i = X^k(\Omega_i; \mathcal{T}_i)$, of local finite element functions that vanish on $\partial\Omega \cap \partial\Omega_i$. In the analysis, we will also employ the GLL mesh $\mathcal{T}_k(\Omega_i)$ on Ω_i, generated by the local GLL meshes $\mathcal{T}_k(K)$ for $K \in \mathcal{T}_i$. The corresponding space of piecewise trilinear functions on $\mathcal{T}_k(\Omega_i)$ that vanish on $\partial\Omega \cap \partial\Omega_i$ is denoted by $Y^h(\Omega_i)$. We set $Y^k(\Omega_i) = X^k(\Omega_i; \mathcal{T}_i)$.

Local bilinear forms We next define the local bilinear forms

$$a_i(u, v) = \int_{\Omega_i} \rho_i \nabla u \cdot \nabla v \, d\mathbf{x}, \quad u, v \in X_i.$$

We note that if Ω_i is a *floating* subdomain (i.e., its boundary does not touch $\partial\Omega$), $a_i(\cdot, \cdot)$ is only positive semi-definite and for $u \in X_i$ we have

$$a_i(u, u) = 0 \quad \text{iff} \quad u \text{ constant in } \Omega_i.$$

The sets of nodal points on Γ_i, Γ, F^{ij}, E^{ij}, and W^i are denoted by $\Gamma_{i,h}$, Γ_h, F_h^{ij}, E_h^{ij}, and W_h^i, respectively. We will identify these sets with the corresponding sets of degrees of freedom. As for the corresponding regions, we will also use notations with one or no superscript.

Extension operators We introduce some spaces defined on the interfaces: U_i is the space of restrictions to Γ_i of functions in $X^k(\Omega_i;\mathcal{T}_i)$ and U of restrictions to Γ of functions in $X^k(\Omega;\mathcal{T})$. We note that functions in U_i and U are uniquely determined by the nodal values in $\Gamma_{i,h}$ and Γ_h, respectively. For every substructure Ω_i, there is a natural interpolation operator

$$R_i^T : U_i \longrightarrow U, \tag{3.25}$$

that extends a function on Γ_i to a global function on Γ with vanishing degrees of freedom in $\Gamma_h \setminus \Gamma_{i,h}$. Its transpose with respect to the Euclidean scalar product $R_i : U \to U_i$ extracts the degrees of freedom in $\Gamma_{i,h}$.

3.5.2. Discrete Harmonic extensions

We introduce next the notion of discrete harmonic extensions and an important lemma that will play a central role.

A function $u^{(i)}$ defined on Ω_i is said to be discrete harmonic on Ω_i if

$$A_{II}^{(i)} u_I^{(i)} + A_{I\Gamma}^{(i)} u_\Gamma^{(i)} = 0.$$

In this case, it is easy to see that $\mathcal{H}_i(u_\Gamma^{(i)}) := u^{(i)}$ is completely defined by its value on Γ_i. The space of piecewise discrete harmonic functions u consists of functions in X that are discrete harmonic on each substructure. In this case, $u =: \mathcal{H}(u_\Gamma)$ is completely defined by its value on Γ.

The balancing domain decomposition preconditioner is defined with respect to the inner product

$$s(u,v) = u^T S v, \quad u, v \in U. \tag{3.26}$$

It follows immediately from the definition of S that $s(\cdot,\cdot)$ is symmetric and coercive.

The following lemma results from elementary variational arguments.

Lemma 3.3. *Let $u_\Gamma^{(i)}$ be the restriction of a finite element function to Γ_i. Then the discrete harmonic extension $u^{(i)} = \mathcal{H}_i(u_\Gamma^{(i)})$ of $u_\Gamma^{(i)}$ into Ω_i satisfies*

$$a_i(u^{(i)}, u^{(i)}) = \min_{v^{(i)}|_{\partial\Omega_i} = u_\Gamma^{(i)}} a_i(v^{(i)}, v^{(i)}) = {u_\Gamma^{(i)}}^T S^{(i)} u_\Gamma^{(i)}.$$

Analogously, if u_Γ is the restriction of a finite element function to Γ, the piecewise discrete harmonic extension $u = \mathcal{H}(u_\Gamma)$ of u_Γ into the interior of the subdomains satisfies

$$a(u,u) = \min_{v|_\Gamma = u_\Gamma} a(v,v) = s(u,u) = u_\Gamma^T S u_\Gamma.$$

3.5.3. Components of the balancing Neumann-Neumann preconditioners

We next define in a variational form the various components of the balancing Neumann-Neumann preconditioner.

P_i projection-like operators

The local operators P_i are projection-like operators associated with a family of subspaces U_i and determined by a set of local bilinear forms defined on them

$$\tilde{s}_i(u,v), \quad u,v \in U_i.$$

Given the interpolation operators $R_i^T : U_i \to U$, we have

$$P_i = R_i^T \tilde{P}_i, \quad \tilde{P}_i : U \longrightarrow U_i, \tag{3.27}$$

with

$$\tilde{s}_i(\tilde{P}_i u, v_i) = s(u, R_i^T v_i), \quad v_i \in U_i. \tag{3.28}$$

For every substructure Ω_i, the local bilinear form is

$$\tilde{s}_i(u,v) := a_i(\mathcal{H}_i(\delta_i u), \mathcal{H}_i(\delta_i v)), \quad u,v \in U_i. \tag{3.29}$$

For a floating subdomain $\tilde{P}_i$ is defined only for those $u \in U$ for which $s(u,v) = 0$ for all $v = R_i^T v_i$ such that $\mathcal{H}_i(\delta_i v_i)$ is constant on Ω_i. This condition is satisfied if $a(u, R_i^T \delta_i^\dagger) = 0$; we note that $R_i^T \delta_i^\dagger$ is a basis function for U_0. For such subdomains, we make the solution $\tilde{P}_i u$ of (3.28) unique by imposing the constraint

$$\int_{\Omega_i} \mathcal{H}_i(\delta_i \tilde{P}_i u) d\mathbf{x} = 0, \tag{3.30}$$

which just means that we select the solution orthogonal to the null space of the Neumann operator.

Coarse space U_0

A coarse space U_0 of minimum dimension is defined as

$$U_0 = \text{span}\{R_i^T \delta_i^\dagger\} \subset U,$$

where the span is taken over the floating subdomains. We note that U_0 consists of piecewise discrete harmonic functions and R_0^T is the natural injection $U_0 \subset U$. We consider an exact solver on U_0

$$\tilde{s}_0(u,v) := a(\mathcal{H}u, \mathcal{H}v) = a(u,v).$$

Partition of unity

An important role in the description and analysis of the Neumann-Neumann algorithms is played by a family of weighted counting functions δ_i, which are associated with and defined on the individual Γ_i (cf. [81, 82, 172, 198, 219]) and are defined for $\gamma \in [1/2, \infty)$. Given Ω_i and $\mathbf{x} \in \Gamma_{i,h}$, $\delta_i(\mathbf{x})$ is determined by a sum of contributions from Ω_i and its relevant nearest neighbours,

$$\delta_i(\mathbf{x}) = \sum_{j \in \mathcal{N}_\mathbf{x}} \rho_j^\gamma(\mathbf{x}) / \rho_i^\gamma(\mathbf{x}), \quad \mathbf{x} \in \Gamma_{i,h}. \tag{3.31}$$

Here $\mathcal{N}_\mathbf{x}$, $\mathbf{x} \in \Gamma_h$, is the set of indices j of the subregions such that $\mathbf{x} \in \Gamma_{j,h}$. The function δ_i is discrete harmonic and thus belongs to U_i. The pseudoinverses $\delta_i^\dagger \in U_i$ are defined, for $\mathbf{x} \in \Gamma_{i,h}$, by

$$\delta_i^\dagger(\mathbf{x}) = \delta_i^{-1}(\mathbf{x}), \quad \mathbf{x} \in \Gamma_{i,h}. \tag{3.32}$$

We note that these functions provide a partition of unity

$$\sum_{i=1}^{M} R_i^T \delta_i{}^\dagger(\mathbf{x}) \equiv 1. \tag{3.33}$$

In particular, for $u \in U$ we can use the formula

$$u = \sum_{i=1}^{M} R_i^T u_i, \qquad \text{with } u_i = \mathcal{H}_i(\delta_i^\dagger u). \tag{3.34}$$

This decomposition result (3.34) is central in the analysis.

3.5.4. Condition number bounds

The main result in this section is a bound for the condition number of the preconditioned operator P_{NN} for *hp* finite element approximations on anisotropic meshes in three dimensions. Such a bound can be found using the abstract Schwarz theory; see, e.g., [226, Ch. 6] or [246, Chapters 2 and 6]. We first derive uniform bounds for the smallest and the largest eigenvalues, successively.

A uniform bound for the smallest eigenvalue is stated next in Lemma 3.4.

Lemma 3.4. *Given U the space of restrictions to Γ of functions in $X^k(\Omega; \mathcal{T})$ and the inner product $s(u, v)$ defined in relation (3.26), the preconditioned operator P_{NN} satisfies the inequality*

$$s(P_{NN}u, u) \geq s(u, u), \quad u \in U.$$

Proof. The result is obtained by using the decomposition (3.34) and the fact that P_0 is an orthogonal projection. □

As stated in, e.g., [226], a stability property for the local bilinear forms is needed in order to find a bound for the largest eigenvalue. This property is stated next in Assumption 3.5.

Assumption 3.5. Given the local interpolation operators R_i^T and projection operators $\tilde{P}_i$, the local bilinear forms $\tilde{s}_i$ satisfy the inequality

$$s(R_i^T u_i, R_i^T u_i) \leq \omega\, \tilde{s}_i(u_i, u_i), \qquad u_i \in \mathcal{R}(\tilde{P}_i), \quad i = 1, \cdots, M,$$

with

$$\omega = C\,(1-\sigma)^{-6}\left(1 + \log\left(\frac{k}{1-\sigma}\right)\right)^2$$

and C independent of k, n, σ, γ, the coefficients ρ_i, and the diameters H_i.

Proof. The fairly technical proof of this assumption is given in [243, Section 9]. These developments extend the proof proposed by Pavarino [198, Section 8] in the context of spectral elements to the case of *hp* finite elements on anisotropic meshes. □

Combining Assumption 3.5 and a colouring argument[1] [246, Section 2.5.1] finally provides a bound for the largest eigenvalue.

Lemma 3.6. *Let Assumption 3.5 be satisfied. Then*

$$s(P_{NN}u, u) \leq C\omega s(u, u), \quad u \in U.$$

Consequently, the condition number of P_{NN} satisfies

$$\kappa(P_{NN}) \leq C\omega = C\,(1-\sigma)^{-6}\left(1 + \log\left(\frac{k}{1-\sigma}\right)\right)^2.$$

Proof. See [243, Sections 7 and 9]. The bound for the condition number can be easily deduced from Lemma 3.4. □

A similar proof has been provided in [242] to derive a bound for the condition number of P_{NN} in two dimensions.

3.6. Additional comments and conclusions

Summary We have proposed two popular iterative substructuring methods (balancing Neumann-Neumann and one-level FETI) for the solution of algebraic systems arising from the *hp* finite element discretization of scalar equations on anisotropic meshes in two and three dimensions. As a main result, we have proved that, if basis functions

[1] Given a decomposition of V, the subspaces $\{V_i, 1 \leq i \leq M\}$ are coloured in such a way that if two subspaces V_k and V_j have the same colour they are orthogonal, i.e., $s(R_k^T u_k, R_j^T u_j) = 0, u_k \in V_k, u_j \in V_j$.

on Gauss-Lobatto-Legendre nodes and the subdomains are suitably chosen, the condition numbers of the preconditioned operator retain the logarithmic dependence in the polynomial degree and they remain independent of arbitrarily large aspect ratios of the mesh and of the number of substructures. *Scalable* preconditioners (in the sense of Definition 1.2) have been thus provided. We have been able to obtain convergence bounds that depend only logarithmically on the spectral polynomial degree k with a constant independent of the possible jumps in the coefficients in the equation. Numerical results in two and three dimensions have been provided supporting the theory.

Collaboration This research on domain decomposition has been made in collaboration with Andrea Toselli with financial support from the Swiss National Science Foundation (SNSF).

Software realization The *hp* finite element method and the domain decomposition preconditioners proposed in this chapter were first implemented in Matlab for fast prototyping in two and three dimensions, respectively. Then, a code was written in Python and C for three-dimensional applications. This code makes extensive use of the Pysparse library[2] written by Roman Geus. Finally, the domain decomposition preconditioners related to vector problems [245] have been implemented in the C++ FEMSTER library [62] for two-dimensional applications. As pointed out in Section 3.1.1, the main emphasis of the research concerns the design and analysis of domain decomposition preconditioners. Parallel implementations of these domain decomposition preconditioners have not been provided but can be derived; see, e.g., discussions in [80, 226] and [241, Section 4].

Short-term prospectives A natural extension of the developments proposed in this chapter would have been to consider dual-primal iterative substructuring methods (BDDC and FETI-DP) methods for *hp* finite element approximations; see, e.g., [105, 159] and the references therein. Among many other advantages, dual-primal FETI algorithms do not require projection steps involving P and P^T, and thus no matrix Q needs to be chosen; see [246, Chapter 6]. Their application to the case of *hp* finite element approximations has been considered in [155]. We also refer the reader to Section 5.3.2 for further discussions on BDDC and FETI-DP methods.

[2] http://pysparse.sourceforge.net/

4. Flexible Krylov subspace methods

4.1. Objectives and contributions

4.1.1. Objectives

In this chapter, we focus on certain Krylov subspace methods for the solution of linear systems of equations

$$Ax = b, \tag{4.1}$$

where $A \in \mathbb{C}^{n\times n}$ is a nonsingular non-Hermitian matrix, $b \in \mathbb{C}^n$, with n assumed to be large. As outlined in Chapter 1, depending on the properties of the PDE that is considered, we assume that a preconditioning method (possibly of multilevel type) is already available. In this chapter, we aim at proposing Krylov subspace methods of minimum residual norm type that combine at least two of the four different features detailed next.

Variable preconditioning

In certain situations, the preconditioning method can be represented by a matrix that changes between the iterations. As already pointed out, this is the case of the multigrid preconditioners that have been proposed in Sections 2.5 and 2.6 for the solution of three-dimensional heterogeneous Helmholtz problems. This also occurs in, e.g., domain decomposition methods, when approximate solvers are considered for the interior problems (see references in [226, Sect. 4.4] or in [246, Sect. 4.3]). This approach is notably used when the size of the local subproblems is so large that obtaining an approximate solution using an iterative method is computationally more interesting than using a direct method. If the domain decomposition preconditioner is based on the use of approximate solvers, its application is not a linear operation in general, and Krylov methods allowing *variable preconditioning* have to be employed (see Section 3.3.4 for an illustration). This class of methods called *flexible methods* [224] will be the central feature of this chapter.

Augmentation

Besides preconditioning, there exist two complementary alternatives to accelerate the convergence of any Krylov subspace method, namely *augmentation* and *deflation*. In augmentation techniques, the search space of the augmented Krylov subspace method is decomposed as a direct sum of two subspaces. This search space S_ℓ (of dimension ℓ) can be written as

$$S_\ell = \mathcal{K}_m(A, b) \oplus \mathcal{W} \tag{4.2}$$

where $\mathcal{K}_m(A, b)$ is the m-th Krylov subspace generated by the matrix A from the vector b

$$\mathcal{K}_m(A, b) = span(b, Ab, \cdots, A^{m-1}b)$$

and $\mathcal{W}$ (of dimension k) is called the augmentation subspace. A typical goal of augmentation is to add information about the problem into the global search space S_ℓ, that is only slowly revealed in the Krylov subspace itself.

Deflation

Alternatively, *deflation* is based on the use of a generic projection operator P. The general idea is to select the projection operator P such that the solution of $\mathsf{P}Ax = \mathsf{P}b$, referred to as the deflated linear system, is more easily amenable to a solution by a Krylov subspace method than the original linear system $Ax = b$. Given the decomposition $x = \mathsf{P}x + (I_n - \mathsf{P})x$, the component $(I_n - \mathsf{P})x$ can then be computed by solving a linear system of small dimension. Usually orthogonal or oblique projection operators with respect to a certain scalar product are employed, depending on the properties of the preconditioned operator. Both choices will be discussed in detail later in this chapter.

Block size reduction

This last feature only concerns the case of a linear system with multiple right-hand sides given simultaneously and written as

$$AX = B, \tag{4.3}$$

where $A \in \mathbb{C}^{n\times n}$ is a nonsingular non-Hermitian matrix, $X, B \in \mathbb{C}^{n\times p}$ with n assumed to be large and $p \leq n$. As emphasised in the literature [136, Section 8], a primary concern when designing efficient block Krylov subspace methods is to remove unneeded information for the convergence as soon as possible during the iterative procedure. This suggests including strategies for detecting when a linear combination of the p systems has approximately converged. This explicit *block size reduction* is often called deflation [136] and should not be confused with deflation introduced above. Hence, in this manuscript, we will rather use the terminology block size reduction. The main purpose is to derive a flexible minimal norm block Krylov subspace method that incorporates block size reduction at each restart or at each iteration suited to the solution of large-scale linear systems (where expensive variable preconditioners are often used) with possibly a large number of right-hand sides. This is especially useful when the cost of the preconditioner is assumed to be larger than the cost of orthogonalization in the block Arnoldi procedure.

4.1.2. Contributions

The main contributions of this chapter are

- FGMRES-DR [124], a flexible Krylov subspace method based on *augmentation* detailed in Section 4.3.3,
- FGCRO-DR [61], a flexible Krylov subspace method based on *augmentation and deflation* detailed in Section 4.3.6 (see Appendix B.3),
- BFGMRES-D [59], a block flexible subspace method including *block size reduction at each restart* detailed in Section 4.5,
- BFGMRES-S [57], a block flexible subspace method including *block size reduction at each iteration* (see Appendix B.5).

4.1.3. Specific notation

Regarding the algorithmic part (Algorithms 4.1-4.9), we adopt Matlab like notation in the presentation. For instance $Q(i, j)$ denotes the q_{ij} entry of matrix Q and $Q(1 : m, 1 : j)$ refers to the submatrix made of the first m rows and first j columns of Q. Given $Z_m = [z_1, \cdots, z_m] \in \mathbb{C}^{n\times m}$ we denote its i-th column as $z_i \in \mathbb{C}^n, (1 \leq i \leq m)$.

4.1.4. Synopsis

We address the case of flexible Krylov subspace methods for the solution of (4.1) in Section 4.3. Then we consider the case of linear systems with multiple right-hand sides (4.3) with block flexible Krylov subspace methods in Section 4.5.

In both sections, we emphasise that we will first derive the mathematical formulation of the numerical methods and then give the corresponding algorithms. Finally, for ease of readability, we briefly recall in Sections 4.2 and 4.4 selected elementary notions related to Krylov subspace methods and block Krylov subspace methods that appear as useful for the developments proposed in this chapter. Conclusions are given in Section 4.6.

4.2. Brief background on Krylov subspace methods

We briefly describe a few basic properties of minimum residual Krylov subspace methods for the solution of (4.1) that will be useful later in this chapter. Finally, throughout Sections 4.2 and 4.3, for the sake of readability, the integer subscript ℓ denotes the dimension of the search space.

4.2.1. Minimum residual Krylov subspace method

In Sections 4.2 and 4.3, we focus on minimum residual norm Krylov subspace methods for the solution of linear systems with a non-Hermitian coefficient matrix. We refer the reader to [217, 257] for a general introduction to Krylov subspace methods and to [224] for a recent overview on Krylov subspace methods; see also [86, 87] for an advanced analysis related to minimum residual norm Krylov subspace methods. Augmented and

deflated minimum residual norm Krylov subspace methods are usually characterized by a generalized Arnoldi relation introduced next.

Definition 4.1. Generalized Arnoldi relation. The minimum residual norm subspace methods investigated in this chapter will satisfy the following relation

$$AZ_\ell = V_{\ell+1}\bar{H}_\ell \tag{4.4}$$

where $Z_\ell \in \mathbb{C}^{n\times\ell}$, $V_{\ell+1} \in \mathbb{C}^{n\times(\ell+1)}$ such that $V_{\ell+1}^H V_{\ell+1} = I_{\ell+1}$ and $\bar{H}_\ell \in \mathbb{C}^{(\ell+1)\times\ell}$. These methods will compute an approximation of the solution of (4.1) in a ℓ-dimensional affine space $x_0 + Z_\ell y_\ell$, where $y_\ell \in \mathbb{C}^\ell$. In certain cases, $\bar{H}_\ell$ will be an upper Hessenberg matrix.

4.2.2. Flexible GMRES

Formulation

Next, we introduce a minimum residual norm subspace method proposed by Saad [215], since it will be the basis for further developments related to augmented and deflated Krylov subspace methods of minimum residual norm type. This method named Flexible GMRES (FGMRES) was primarily introduced to allow variable preconditioning. We denote by M_j the nonsingular matrix that represents the preconditioner at step j of the method. Starting from an initial guess $x_0 \in \mathbb{C}^n$, it is based on a generalized Arnoldi relation (4.4), where $Z_\ell \in \mathbb{C}^{n\times\ell}$, $V_{\ell+1} \in \mathbb{C}^{n\times(\ell+1)}$ and the upper Hessenberg matrix $\bar{H}_\ell \in \mathbb{C}^{(\ell+1)\times\ell}$ are obtained from the Arnoldi procedure described in Algorithm 4.1. An approximate solution $x_\ell \in \mathbb{C}^n$ is then found by minimizing the residual norm $\|b - A(x_0 + Z_\ell y)\|_2$ over the space $x_0 + \mathcal{R}(Z_\ell)$, the corresponding residual being $r_\ell = b - Ax_\ell \in \mathbb{C}^n$ with $r_\ell \in \mathcal{R}(V_{\ell+1})$. The current approximation x_ℓ can be written as

$$x_\ell = x_0 + Z_\ell y^\star, \tag{4.5}$$

whereas the residual $r_\ell = b - Ax_\ell$ satisfies the Petrov-Galerkin orthogonality condition

$$r_\ell \perp A\ \mathcal{R}(Z_\ell).$$

Hence, an optimality property similar to the one that defines GMRES is thus obtained [217]. We note however that no general convergence results are available since the subspace of approximants $\mathcal{R}(Z_\ell)$ is no longer a standard Krylov subspace. We refer the reader to [215, 217] for the analysis of breakdown in FGMRES. Furthermore, as can be seen in equation (4.5), the update of the iterate x_ℓ requires storing the complete set of vectors Z_ℓ causing a possibly large memory footprint for large ℓ. In order to alleviate this memory requirement, a restarting strategy must be implemented as shown in Algorithm 4.2. The construction of a complete set of Z_ℓ is often named a cycle of the method: it corresponds to one iteration of the loop in Algorithm 4.2. When the preconditioner is fixed, FGMRES(ℓ) reduces to right-preconditioned GMRES(ℓ), whose convergence properties are discussed in, e.g., [217, Chapter 6].

Algorithms

Arnoldi procedure Algorithm 4.1 introduces the Arnoldi procedure with modified Gram-Schmidt. Algorithm 4.1 proceeds by orthonormalizing Az_j against all the previous preconditioned Krylov directions.

Algorithm 4.1 Arnoldi procedure: computation of $V_{\ell+1}$, Z_ℓ and $\bar{H}_\ell$

Input: Assume that the following is given:

- $A \in \mathbb{C}^{n\times n}$ ▷ operator
- $v_1 \in \mathbb{C}^n$ ▷ vector of unit Euclidean norm ($\|v_1\|_2 = 1$)
- $M_j^{-1} \in \mathbb{C}^{n\times n}$ for $j \in \{1,\dots,\ell\}$ ▷ variable preconditioning operators

1: **for** $j = 1, \ell$ **do**
2: $z_j = M_j^{-1} v_j$
3: $s = Az_j$
4: **for** $i = 1, j$ **do**
5: $h_{i,j} = v_i^H s$
6: $s = s - h_{i,j} v_i$
7: **end for**
8: $h_{i+1,j} = \|s\|_2$
9: $v_{j+1} = s/h_{i+1,j}$
10: **end for**
11: Define $Z_\ell = [z_1, \cdots, z_\ell]$, $V_{\ell+1} = [v_1, \cdots, v_{\ell+1}]$, $\bar{H}_\ell = \{h_{i,j}\}_{1\le i\le \ell+1, 1\le j\le \ell}$

FGMRES(ℓ) Algorithm 4.2 depicts the FGMRES(ℓ) method, where the dimension of the approximation subspace is not allowed to be larger than a prescribed dimension noted ℓ.

4.3. Flexible augmented and deflated Krylov subspace methods

4.3.1. Problem setting

In this section, we examine augmentation and deflation techniques in Krylov subspace methods when the coefficient matrix A is non-Hermitian. We specifically focus on minimum residual norm subspace methods and assume that a generalized Arnoldi relation (4.4) holds. We denote by x_0, $r_0 = b - Ax_0$ the current approximation and residual vector respectively, and by $V_{\ell+1}$, $\bar{H}_\ell$ and Z_ℓ the matrices involved in this relation. With notation of Algorithm 4.2, r_0 can be expressed as $r_0 = V_{\ell+1}(c - \bar{H}_\ell y^\star)$.

Algorithm 4.2 Flexible GMRES(ℓ)

Input: Assume that the following is given:

- $A \in \mathbb{C}^{n\times n}$ ▷ operator
- $b, x_0 \in \mathbb{C}^n$ ▷ right-hand side and initial guess
- $M_j^{-1} \in \mathbb{C}^{n\times n}$ for $j \in \{1, \ldots, \ell\}$ ▷ variable preconditioning operators
- $cycle_{\max}$ ▷ maximal number of cycles allowed
- $tol > 0$ ▷ convergence threshold

1: *Settings:* Let $r_0 = b - Ax_0$, $\beta = \|r_0\|_2$, $c = [\beta, 0_{1\times\ell}]^T$ where $c \in \mathbb{C}^{\ell+1}$, $v_1 = r_0/\beta$.
2: **for** $cycle = 1, cycle_{\max}$ **do**
3: *Computation of $V_{\ell+1}$, Z_ℓ and $\bar{H}_\ell$ (see Algorithm 4.1):* Apply ℓ steps of the Arnoldi method with variable preconditioning ($z_j = M_j^{-1}v_j, 1 \le j \le \ell$) to obtain $V_{\ell+1} \in \mathbb{C}^{n\times(\ell+1)}$, $Z_\ell \in \mathbb{C}^{n\times\ell}$ and the upper Hessenberg matrix $\bar{H}_\ell \in \mathbb{C}^{(\ell+1)\times\ell}$ such that

$$AZ_\ell = V_{\ell+1}\bar{H}_\ell \quad \text{with} \quad V_{\ell+1}^H V_{\ell+1} = I_{\ell+1}.$$

4: *Minimum norm solution:* Compute the minimum norm solution $x_\ell \in \mathbb{C}^n$ in the affine space $x_0 + \mathcal{R}(Z_\ell)$; that is, $x_\ell = x_0 + Z_\ell y^\star$ where $y^\star = \operatorname{argmin}_{y\in\mathbb{C}^\ell} \|c - \bar{H}_\ell y\|_2$.
5: *Check the convergence criterion:* If $\|c - \bar{H}_\ell y^\star\|_2/\|b\|_2 \le tol$, exit
6: *Settings:* Set $x_0 = x_\ell$, $r_0 = b - Ax_0$, $\beta = \|r_0\|_2$, $c = [\beta, 0_{1\times\ell}]^T$, $v_1 = r_0/\beta$.
7: **end for**

4.3.2. Augmented Krylov subspace methods

We next discuss two possibilities for selecting the augmentation subspace $\mathcal{W}$ and analysing the corresponding Krylov subspace methods.

Augmentation with an arbitrary subspace

Given an augmentation subspace $\mathcal{W}$ (subspace of $\mathbb{C}^n$ of dimension k) with $W = [w_1, \cdots, w_k]$ a matrix whose columns form a basis of $\mathcal{W}$, a slight modification in the Arnoldi procedure (Algorithm 4.1) is used to obtain an orthogonal basis of S_ℓ defined in (4.2) (see [63]). It consists of defining z_j (line 2 of Algorithm 4.1) now as

$$z_j = M_j^{-1}v_j \quad (1 \le j \le m), \quad z_j = M_j^{-1}w_{j-m} \quad (m < j \le m + k).$$

With this definition, we finally obtain the generalized Arnoldi relation

$$AZ_{m+k} = V_{m+k+1}\bar{H}_{m+k}$$

where

$$Z_{m+k} = [M_1^{-1}v_1, M_2^{-1}v_2, \cdots, M_{m+1}^{-1}w_1, M_{m+2}^{-1}w_2, \cdots, M_{m+k}^{-1}w_k], \tag{4.6}$$
$$V_{m+k+1} = [v_1, v_2, \cdots, v_{m+k+1}], \tag{4.7}$$

and $\bar{H}_{m+k}$ is a $(m+k+1) \times (m+k)$ upper Hessenberg matrix. Thus, the residual minimization properties are then obtained similarly to FGMRES [215]. Hence, the approximate solution from the affine space $x_0 + \mathcal{R}(Z_{m+k})$ can be written as

$$x_{m+k} = x_0 + Z_{m+k} y^\star$$

with $y^\star \in \mathbb{C}^{(m+k)}$ the solution of the residual norm minimization problem

$$y^\star = \text{argmin}_{y \in \mathbb{C}^{(m+k)}} \| \|r_0\| e_1 - \bar{H}_{m+k} y \|_2,$$

(with e_1 designating here the first unit vector of $\mathbb{R}^{(m+k+1)}$). In the case of fixed right preconditioning, the main important property is that if any vector w_j, $1 \leq k$, is the solution of $AM^{-1}w_j = v_i, 1 \leq i \leq m$, then in general the exact solution of the original system (4.1) can be extracted from S_ℓ; see, e.g., [216, Proposition 2.1]. We refer the reader to [63] for a discussion of possible choices for the augmented subspace $\mathcal{W}$. Vectors obtained with either different iterative methods or with different preconditioners can be incorporated into Z_{m+k} quite easily. A popular idea is to choose $\mathcal{W}$ as an approximate invariant subspace associated with a specific part of the spectrum of A or AM^{-1} in the case of fixed preconditioning. This is discussed next.

Augmentation with an approximate invariant subspace

A typical goal of augmentation is to add information about the problem into the search space that is only slowly revealed in the Krylov subspace itself. In the symmetric positive definite case, it is often known that eigenvalues of the (preconditioned) operator close to zero tend to slow down the convergence rate of the Krylov subspace methods [63]. Hence, augmentation based on approximate invariant subspaces made of eigenvectors corresponding to eigenvalues small in modulus of the (preconditioned) operator has been proposed; see, e.g., [183, 184, 185, 216] and references therein.

Harmonic Ritz information In [184], Morgan has suggested selecting $\mathcal{W}$ as an approximate invariant subspace and updating this subspace at the end of each cycle. Approximate spectral information is then required to define the augmentation space. This is usually obtained by computing harmonic Ritz pairs of A with respect to a certain subspace [63, 184]. We present here a definition of a harmonic Ritz pair as given in [196, 225].

Definition 4.2. Harmonic Ritz pair. Consider a subspace $\mathcal{U}$ of $\mathbb{C}^n$. Given $B \in \mathbb{C}^{n \times n}$, $\theta \in \mathbb{C}$ and $y \in \mathcal{U}$, (θ, y) is a harmonic Ritz pair of B with respect to $\mathcal{U}$ if and only if

$$By - \theta y \perp B\mathcal{U}$$

or equivalently, for the canonical scalar product,

$$\forall w \in \mathcal{R}(B\mathcal{U}) \quad w^H (By - \theta y) = 0.$$

We call y a harmonic Ritz vector associated with the harmonic Ritz value θ.

Based on the generalized Arnoldi relation (4.4), the augmentation procedure proposed in [124, Proposition 1] relies on the use of k harmonic Ritz vectors $Y_k = V_\ell P_k$ of $AZ_\ell V_\ell^H$ with respect to $\mathcal{R}(V_\ell)$, where $Y_k \in \mathbb{C}^{n\times k}$ and $P_k = [p_1, \cdots, p_k] \in \mathbb{C}^{\ell\times k}$. According to Definition 4.2, the harmonic Ritz vector $y_j = V_\ell p_j$ then satisfies

$$Z_\ell^H A^H \left(AZ_\ell\ p_j - \theta_j V_\ell\ p_j\right) \quad = \quad 0. \tag{4.8}$$

Using the generalized Arnoldi relation (4.4) we finally obtain the relation

$$\bar{H}_\ell^H\ \bar{H}_\ell\ y_j = \theta \bar{H}_\ell^H V_{\ell+1}^H V_\ell\ y_j. \tag{4.9}$$

Since

$$\bar{H}_\ell = \begin{bmatrix} H_\ell \\ h_{\ell+1,\ell} e_\ell^T \end{bmatrix}, \quad H_\ell \in \mathbb{C}^{\ell\times\ell}$$

where $H_\ell \in \mathbb{C}^{\ell\times\ell}$ is assumed to be nonsingular, the generalized eigenvalue problem is then equivalent to

$$(H_\ell + h_{\ell+1,\ell}^2 H_\ell^{-H} e_\ell e_\ell^T) y_j = \theta_j y_j. \tag{4.10}$$

This corresponds to a standard eigenvalue problem of dimension ℓ only, where ℓ is assumed to be much smaller than the problem dimension n. As a consequence, the approximate spectral information based on Harmonic Ritz pair is quite inexpensive to compute.

4.3.3. Flexible GMRES with deflated restarting: FGMRES-DR

We present next a flexible Krylov subspace method based on *augmentation*, which consists of our first contribution of this chapter. We introduce the mathematical derivation of the method in Section 4.3.3, and follow it by an algorithmic description given in Section 4.3.3.

Motivations

Formulation

The augmentation space $\mathcal{W}$ based on approximate invariant information corresponding to $\mathcal{R}(Y_k)$ is used. The key point studied next is to understand how to incorporate this information into a minimum residual norm subspace method such as flexible GMRES. This has been proposed in [124]. The i-th cycle of the resulting algorithm called FGMRES-DR is now briefly described, and we denote by $r_0^{(i-1)} = b - Ax_0^{(i-1)}$, $V_{\ell+1}$, $\bar{H}_\ell$ and Z_ℓ the residual and matrices obtained at the end of the $(i-1)$-th cycle. Based on the generalized Arnoldi relation (4.4), the augmentation procedure proposed in [124, Proposition 1] relies on the use of k harmonic Ritz vectors $Y_k = V_\ell P_k$ of $AZ_\ell V_\ell^H$ with respect to $\mathcal{R}(V_\ell)$, where $Y_k \in \mathbb{C}^{n\times k}$ and $P_k \in \mathbb{C}^{\ell\times k}$. In Lemma 4.3 shown in [124, Lemma 3.1], we recall a useful relation satisfied by the harmonic Ritz vectors.

Lemma 4.3. *In flexible GMRES with deflated restarting, the harmonic Ritz vectors are given by* $Y_k = V_\ell P_k$ *with corresponding harmonic Ritz values* λ_k. $P_k \in \mathbb{C}^{\ell\times k}$ *satisfies the relation*

$$AZ_\ell P_k = V_{\ell+1}\left[\begin{bmatrix} P_k \\ 0_{1\times k}\end{bmatrix}, c - \bar{H}_\ell y^*\right]\begin{bmatrix} \operatorname{diag}(\lambda_1,\ldots,\lambda_k) \\ \alpha_{1\times k}\end{bmatrix}, \tag{4.11}$$

$$AZ_\ell P_k = [V_\ell P_k, r_0^{(i-1)}]\begin{bmatrix} \operatorname{diag}(\lambda_1,\ldots,\lambda_k) \\ \alpha_{1\times k}\end{bmatrix}, \tag{4.12}$$

where $r_0^{(i-1)} = V_{\ell+1}(c - \bar{H}_\ell y^*)$ *and* $\alpha_{1\times k} = [\alpha_1,\ldots,\alpha_k] \in \mathbb{C}^{1\times k}$.

Proof. See Lemma 3.1 of [124]. □

Next, the QR factorization of the $(\ell+1)\times(k+1)$ matrix appearing on the right-hand side of relation (4.11) is performed as

$$\left[\begin{bmatrix} P_k \\ 0_{1\times k}\end{bmatrix}, c - \bar{H}_\ell y^*\right] = QR \tag{4.13}$$

where $Q \in \mathbb{C}^{(\ell+1)\times(k+1)}$ has orthonormal columns and $R \in \mathbb{C}^{(k+1)\times(k+1)}$ is upper triangular, respectively. We write the matrix Q in equation (4.13) as

$$Q = \left[\begin{bmatrix} Q_{\ell\times k} \\ 0_{1\times k}\end{bmatrix}, \frac{\bar{\rho}}{\|\bar{\rho}\|}\right], \tag{4.14}$$

where $Q_{\ell\times k} \in \mathbb{C}^{\ell\times k}$ and $\bar{\rho} \in \mathbb{C}^{\ell+1}$ is defined as

$$\bar{\rho} = (I_{\ell+1} - \begin{bmatrix} Q_{\ell\times k} \\ 0_{1\times k}\end{bmatrix}\begin{bmatrix} Q_{\ell\times k} \\ 0_{1\times k}\end{bmatrix}^H)(c - \bar{H}_\ell y^*). \tag{4.15}$$

Proposition 4.4. *In flexible GMRES with deflated restarting, the generalized Arnoldi relation*

$$A\ Z_k = V_{k+1}\bar{H}_k, \tag{4.16}$$

$$V_{k+1}^H V_{k+1} = I_{k+1}, \tag{4.17}$$

$$\mathcal{R}([Y_k, r_0^{(i-1)}]) = \mathcal{R}(V_{k+1}) \tag{4.18}$$

holds at the i*-th cycle with matrices* $Z_k, V_k \in \mathbb{C}^{n\times k}$ *and* $\bar{H}_k \in \mathbb{C}^{(k+1)\times k}$ *defined as*

$$Z_k = Z_\ell Q_{\ell\times k}, \tag{4.19}$$

$$V_{k+1} = V_{\ell+1}Q, \tag{4.20}$$

$$\bar{H}_k = Q^H \bar{H}_\ell Q_{\ell\times k}, \tag{4.21}$$

where $V_{\ell+1}$, Z_ℓ *and* $\bar{H}_\ell$ *refer to matrices obtained at the end of the* $(i-1)$*-th cycle.*

Proof. Relations (4.16), (4.17), (4.19), (4.20) and (4.21) have been shown in [124, Proposition 2]. From relations (4.20) and (4.13) respectively, we deduce that

$$\begin{aligned} V_{k+1}R &= V_{\ell+1}\left[\begin{bmatrix} P_k \\ 0_{1\times k}\end{bmatrix}, c - \bar{H}_\ell y^*\right] \\ V_{k+1}R &= [V_\ell P_k, r_0^{(i-1)}] \end{aligned} \tag{4.22}$$

which finally shows that $\mathcal{R}([Y_k, r_0^{(i-1)}]) = \mathcal{R}(V_{k+1})$ since R is assumed to be nonsingular. □

FGMRES-DR then carries out m Arnoldi steps with variable preconditioning and starting vector v_{k+1}, while maintaining orthogonality to V_k leading to

$$A\,[z_{k+1}, \cdots, z_{m+k}] = [v_{k+1}, \cdots, v_{m+k+1}]\,\bar{H}_m \quad \text{and} \quad V_{m+k+1}^H\,V_{m+k+1} = I_{m+k+1}.$$

We note that $\bar{H}_m \in \mathbb{C}^{(m+1)\times(m)}$ is upper Hessenberg. At the end of the i-th cycle this gives the generalized Arnoldi relation

$$A\,Z_{m+k} = [V_{m+1+k}]\left[\begin{bmatrix} \bar{H}_k \\ 0_{m\times k}\end{bmatrix}\begin{bmatrix} B_{k\times m} \\ \bar{H}_m\end{bmatrix}\right]$$

where $V_{m+k+1} \in \mathbb{C}^{n\times(m+k+1)}$, $\bar{H}_{m+k} \in \mathbb{C}^{(m+k+1)\times(m+k)}$ and $B_{k\times m} \in \mathbb{C}^{k\times m}$ results from the orthogonalization of $[v_{k+2}, \cdots, v_{m+k+1}]$ against V_{k+1}. We note that $\bar{H}_{m+k}$ is no longer upper Hessenberg due to the leading dense $(k+1)\times k$ submatrix $\bar{H}_k$. At the end of the i-th cycle, an approximate solution $x_0^{(i)} \in \mathbb{C}^n$ is then found by minimizing the residual norm $\|b - A(x_0^{(i-1)} + Z_\ell y)\|_2$ over the space $x_0^{(i-1)} + \mathcal{R}(Z_\ell)$, the corresponding residual being $r_0^{(i)} = b - Ax_0^{(i)}$, with $r_0^{(i)} \in \mathcal{R}(V_{\ell+1})$.

Algorithms

FGMRES-DR is depicted in Algorithms 4.3 and 4.4, respectively. We refer the reader to [124, Section 3.2] for a complete description of the implementation and details on related computational aspects.

Remarks When the preconditioner is fixed, the previous algorithm is known as GMRES with deflated restarting (GMRES-DR), initially proposed by Morgan [184]. Although the term “deflated" is used, we note that this algorithm does correspond to a GMRES method with an adaptive augmented basis without any explicit deflated matrix. The success of GMRES-DR has been demonstrated on many academic examples [183] and concrete applications, such as in lattice QCD [72, 110], reservoir modelling [6, 160] or electromagnetics [124]. We refer the reader to [184, 210] for further comments on the algorithm and computational details.

Finally, we refer the reader to [124] for various numerical experiments showing the efficiency of FGMRES-DR on both academic and industrial applications.

Algorithm 4.3 Flexible GMRES with deflated restarting: FGMRES-DR(m, k).

Input: Assume that the following is given:

- $A \in \mathbb{C}^{n\times n}$ ▷ operator
- $b \in \mathbb{C}^n, x_0 \in \mathbb{C}^n$ ▷ right-hand side and initial guess
- $M_j^{-1} \in \mathbb{C}^{n\times n}$ for $j \in \{1, \ldots, \ell\}$ ▷ variable preconditioning operators
- $cycle_{\max}$ ▷ maximal number of cycles allowed
- $tol > 0$ ▷ convergence threshold

1: *Settings:* Let $r_0 = b - Ax_0$, $\beta = \|r_0\|_2$, $c = [\beta, 0_{1\times m}]^T \in \mathbb{C}^{m+1}$, $v_1 = r_0/\beta$.
2: *Computation of V_{m+1}, Z_m, and $\bar{H}_m$:* Apply m steps of the Arnoldi procedure with *flexible* preconditioning to obtain $V_{m+1} \in \mathbb{C}^{n\times(m+1)}$, $Z_m \in \mathbb{C}^{n\times m}$, and the upper Hessenberg matrix $\bar{H}_m \in \mathbb{C}^{(m+1)\times m}$ such that

$$AZ_m = V_{m+1}\bar{H}_m \quad \text{with} \quad V_{m+1}^H V_{m+1} = I_{m+1}.$$

3: **for** $cycle = 1, cycle_{\max}$ **do**
4: *Minimum norm solution:* Compute the minimum norm solution $x_m \in \mathbb{C}^n$ in the affine space $x_0 + \mathcal{R}(Z_m)$; that is, $x_m = x_0 + Z_m y^*$, where $y^* = \operatorname{argmin}_{y\in\mathbb{C}^m} \|c - \bar{H}_m y\|_2$. Set $x_0 = x_m$ and $r_0 = b - Ax_0$.
5: *Check the convergence criterion:* If $\|c - \bar{H}_m y^*\|_2/\|b\|_2 \le tol$, exit.
6: *Computation of V_{k+1}^{new}, Z_k^{new}, and $\bar{H}_k^{new}$:* see Algorithm 4.4. At the end of this step the following relations hold:

$$AZ_k^{new} = V_{k+1}^{new}\bar{H}_k^{new} \text{ with } V_{k+1}^{new\,H}V_{k+1}^{new} = I_{k+1} \text{ and } r_0 \in \mathcal{R}(V_{k+1}^{new}). \tag{4.23}$$

7: *Arnoldi procedure:* Set $V_{k+1} = V_{k+1}^{new}$, $Z_k = Z_k^{new}$, and $\bar{H}_k = \bar{H}_k^{new}$ and apply $(m-k)$ steps of the Arnoldi procedure with *flexible* preconditioning and starting vector v_{k+1} to build $V_{m+1} \in \mathbb{C}^{n\times(m+1)}$, $Z_m \in \mathbb{C}^{n\times m}$, and $\bar{H}_m \in \mathbb{C}^{(m+1)\times m}$ such that

$$AZ_m = V_{m+1}\bar{H}_m \quad \text{with} \quad V_{m+1}^H V_{m+1} = I_{m+1}.$$

8: *Setting:* Set $c = V_{m+1}^H r_0$.
9: **end for**

4.3.4. Deflated Krylov subspace methods

We next briefly describe minimal residual Krylov subspace methods based on deflation. We refer the reader to [121, 122, 138] for a recent excellent overview of deflated Krylov subspace methods in the Hermitian and non-Hermitian case, where extensive bibliographical references and historical comments can be found. The general idea of deflation is to decompose the approximation space into two complementary subspaces such that the projected linear system, referred to as the deflated linear system, will be easier to solve iteratively than the original linear system (4.1). The fact that these subspaces can be chosen in different ways explains the existence of a huge literature on deflated Krylov subspace methods. The Krylov subspace method is then confined to

Algorithm 4.4 FGMRES-DR(m,k): computation of V_{k+1}^{new}, Z_k^{new}, and $\bar{H}_k^{new}$.

Input: Assume that the following is given:

- $A \in \mathbb{C}^{n\times n}$ ▷ operator
- Z_m, V_{m+1} such that $AZ_m = V_{m+1}\bar{H}_m$ ▷ generalized Arnoldi relation
- $c - \bar{H}_m y^*$ ▷ $r_0 = V_{m+1}(c - \bar{H}_m y^*)$

1: *Settings:* Define $h_{m+1,m} = \bar{H}_m(m+1,m)$, $H_m \in \mathbb{C}^{m\times m}$ as $H_m = \bar{H}_m(1:m,1:m)$.
2: *Compute k harmonic Ritz vectors.* Compute k independent eigenvectors g_i of the matrix $H_m + |h_{m+1,m}|^2 H_m^{-H} e_m e_m^T$. Set $G_k = [g_1,\ldots,g_k] \in \mathbb{C}^{m\times k}$.
3: *Augmentation of G_k:* Define $G_{k+1} \in \mathbb{C}^{(m+1)\times(k+1)}$ as

$$G_{k+1} = \begin{bmatrix} \begin{bmatrix} G_k \\ 0_{1\times k} \end{bmatrix}, c - \bar{H}_m y^* \end{bmatrix}. \tag{4.24}$$

4: *Orthonormalization of the columns of G_{k+1}:* Perform a QR factorization of G_{k+1} as $G_{k+1} = P_{k+1}\Gamma_{k+1}$. Define $P_k \in \mathbb{C}^{m\times k}$ as $P_k = P_{k+1}(1:m,1:k)$.
5: *Settings and final relation:* Set $V_{k+1}^{new} = V_{m+1}P_{k+1}$, $Z_k^{new} = Z_m P_k$, and $\bar{H}_k^{new} = P_{k+1}^H \bar{H}_m P_k$, so that the following relations are satisfied:

$$AZ_m P_k = V_{m+1}P_{k+1}P_{k+1}^H \bar{H}_m P_k; \quad \text{i.e.,} \quad AZ_k^{new} = V_{k+1}^{new}\bar{H}_k^{new}, \tag{4.25}$$

where $\bar{H}_k^{new}$ is generally a dense matrix.

one of these subspaces, by projecting the initial residual into this space and by replacing A by its restriction to this subspace. If the projection operator is chosen properly the deflated linear system will be easier to solve iteratively than the original linear system (4.1). We first present a possible strategy based on orthogonal projection and then briefly discuss an extension based on oblique projection proposed in [138].

Deflation based on orthogonal projection

We still denote by $\mathcal{W}$ a subspace of $\mathbb{C}^n$ of dimension k, where k is assumed to be much smaller than the problem dimension n. We later denote by $W \in \mathbb{C}^{n\times k}$, a matrix whose columns form a basis of $\mathcal{W}$ so that $W^H A^H A W$ is Hermitian positive definite (hence invertible). To simplify further developments, we introduce the operators $\mathsf{P}^{\mathrm{MR}}_{(\mathsf{A}\mathcal{W})^\perp}, \mathsf{P}^{\mathrm{MR}}_{\mathcal{W}^\perp} \in \mathbb{C}^{n\times n}$ defined respectively as

$$\mathsf{P}^{\mathrm{MR}}_{(\mathsf{A}\mathcal{W})^\perp} = I_n - AW(W^H A^H A W)^{-1} W^H A^H, \tag{4.26}$$

$$\mathsf{P}^{\mathrm{MR}}_{\mathcal{W}^\perp} = I_n - W(W^H A^H A W)^{-1} W^H A^H A. \tag{4.27}$$

We can easily show that $\mathsf{P}^{\mathrm{MR}}_{(\mathsf{A}\mathcal{W})^\perp}$ and $\mathsf{P}^{\mathrm{MR}}_{\mathcal{W}^\perp}$ are orthogonal projectors such that $\mathsf{P}^{\mathrm{MR}}_{(\mathsf{A}\mathcal{W})^\perp}$ projects onto $(A\mathcal{W})^\perp$, whereas $\mathsf{P}^{\mathrm{MR}}_{\mathcal{W}^\perp}$ projects onto $\mathcal{W}^\perp$, both with respect to the inner product $\langle\cdot,\cdot\rangle_{\mathsf{A}^H\mathsf{A}}$. Furthermore, we note that $\mathsf{P}^{\mathrm{MR}}_{(\mathsf{A}\mathcal{W})^\perp}$ is Hermitian and that $A\mathsf{P}^{\mathrm{MR}}_{\mathcal{W}^\perp} =$

$\mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}A$. The decomposition based on orthogonal projection reads as

$$\mathbb{C}^n = \mathcal{W} \oplus \mathcal{W}^\perp. \tag{4.28}$$

Hence, the solution x of the original system (4.1) can be written as

$$x = (I_n - \mathsf{P}^{\mathrm{MR}}_{\mathcal{W}^\perp})x + \mathsf{P}^{\mathrm{MR}}_{\mathcal{W}^\perp}x = W(W^HA^HAW)^{-1}W^HA^Hb + \mathsf{P}^{\mathrm{MR}}_{\mathcal{W}^\perp}x.$$

Combining this decomposition with $A\mathsf{P}^{\mathrm{MR}}_{\mathcal{W}^\perp} = \mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}A$, the original system (4.1) then simply becomes

$$\mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}Ax = \mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}b. \tag{4.29}$$

Since $\mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}AW = 0_{n\times k}$, $\mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}A$ is singular. Although this deflated matrix $\mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}A$ is singular, the deflated linear system (4.29) is consistent and thus can be solved by an appropriate Krylov subspace method. Here, we focus on the application of a minimum residual Krylov subspace method based on GMRES to solve the deflated linear system (4.29). Hence, given an initial guess x_0 and initial residual $r_0 = b - Ax_0$, the search space of the Krylov subspace method applied to (4.29) can be written as

$$\hat{S}_m = \mathcal{K}_m(\mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}A, \mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}r_0),$$

while the current approximation $\hat{x}_m$ and the current residual $\hat{r}_m = \mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}(b - A\hat{x}_m)$ at the end of the cycle satisfy the relations

$$\begin{aligned} \hat{x}_m &\in \hat{x}_0 + \hat{S}_m, \\ \mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}(b - A\hat{x}_m) &\perp \mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}A\ \mathcal{K}_m(\mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}A, \mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}r_0). \end{aligned}$$

Hence, it is of paramount importance to analyse the possibilities of a breakdown when solving the deflated linear system (4.29). In our context, when GMRES is used to solve the deflated linear system, this feature has been notably analysed in [138, Section 3] based on theoretical results obtained by Brown and Walker [46]. We refer the reader to [138, Corollary 3] for conditions that characterize the possibility of breakdowns. It is worthwhile to note that a breakdown cannot occur if the condition

$$\mathcal{N}(\mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}A) \cap \mathcal{R}(\mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}A) = \{0\}$$

holds; see [122, Theorem 4.1]. This condition is notably satisfied if $\mathcal{W}$ is chosen as an exact A-invariant subspace, i.e., when $A\mathcal{W} = \mathcal{W}D$ since $\mathcal{N}(\mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}A) = \mathcal{W}$ and $\mathcal{R}(\mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}A) = \mathcal{W}^\perp$, due to the nonsingularity of A. Once the solution of the deflated linear system (4.29) is obtained, we deduce the approximation x_m of the original system as

$$x_m = W(W^HA^HAW)^{-1}W^HA^Hb + \mathsf{P}^{\mathrm{MR}}_{\mathcal{W}^\perp}\hat{x}_m,$$

and by construction we note that

$$b - Ax_m = \mathsf{P}^{\mathrm{MR}}_{(A\mathcal{W})^\perp}(b - A\hat{x}_m),$$

i.e.,

$$r_m = \hat{r}_m.$$

We refer to [97] for applications of deflated Krylov subspace methods with orthogonal projection to linear systems with non-Hermitian matrices. As an illustration, a typical choice of subspaces is to choose the columns of W as right eigenvectors of A corresponding to eigenvalues of small modulus.

Deflation based on oblique projection

We briefly mention a strategy based on the use of an oblique projection that is considered as more appropriate for the solution of non-Hermitian linear systems, since the eigenspaces of A are in general not mutually orthogonal [138]. As in (4.28), the search space S_ℓ will be decomposed into a direct sum of two subspaces. More precisely, the following decompositions into nonorthogonal complements are used

$$\mathbb{C}^n = A\mathcal{W} \oplus \tilde{\mathcal{W}}^\perp = A\tilde{\mathcal{W}} \oplus \mathcal{W}^\perp,$$

where $\mathcal{W}$ and $\tilde{\mathcal{W}}$ represent two subspaces of $\mathbb{C}^n$ of dimension k, respectively. As before, we denote by $W \in \mathbb{C}^{n\times k}$ ($\tilde{W} \in \mathbb{C}^{n\times k}$) a matrix whose columns form a basis of $\mathcal{W}$ ($\tilde{\mathcal{W}}$, respectively). We assume that both matrices are chosen such that $\tilde{W}^H AW$ is nonsingular. The key idea is then to introduce the matrices $\mathsf{P}_{A\mathcal{W},\tilde{\mathcal{W}}^\perp} \in \mathbb{C}^{n\times n}, \mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}} \in \mathbb{C}^{n\times n}$ defined as

$$\mathsf{P}_{A\mathcal{W},\tilde{\mathcal{W}}^\perp} = AW(\tilde{W}^H AW)^{-1}\tilde{W}^H, \tag{4.30}$$

$$\mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}} = I_n - AW(\tilde{W}^H AW)^{-1}\tilde{W}^H. \tag{4.31}$$

It is easy to show that $\mathsf{P}_{A\mathcal{W},\tilde{\mathcal{W}}^\perp}$ and $\mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}} = I_n - \mathsf{P}_{A\mathcal{W},\tilde{\mathcal{W}}^\perp}$ are projection operators; $\mathsf{P}_{A\mathcal{W},\tilde{\mathcal{W}}^\perp}$ is the oblique projection onto $(A\mathcal{W})$ along $\tilde{\mathcal{W}}^\perp$, while $\mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}}$ is the oblique projection onto $\tilde{\mathcal{W}}^\perp$ along $A\mathcal{W}$ with respect to the (possibly semidefinite) inner product $\langle\cdot,\cdot\rangle_{\mathsf{A}}$. Given these oblique projection operators, the deflated linear system is now defined as

$$\mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}} A \mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}} x = \mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}} b$$

with $\hat{r}_0 = \mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}} r_0$. The use of a Krylov subspace method is now restricted to $\tilde{\mathcal{W}}^\perp$. Hence, it can be shown that the deflated Krylov subspace method based on GMRES yields iterates x_m at the end of the cycle of the form

$$x_m \in x_0 + \mathcal{K}_m(\mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}} A \mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}}, \mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}} r_0) + \mathcal{W}.$$

This also implies the following relation for the residual [138]

$$b - Ax_m \in r_0 + A\mathcal{K}_m(\mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}} A \mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}}, \mathsf{P}_{\tilde{\mathcal{W}}^\perp,A\mathcal{W}} r_0) + A\mathcal{W}.$$

We refer the reader to [138, Sections 5 and 6] for the mathematical aspects of deflated Krylov subspace methods based on oblique projections and to [138, Section 11] for

an overview of partly related methods that only differ in the choice of the projection operators. A typical choice is to choose the columns of W as right eigenvectors of A and the columns of $\tilde{W}$ as the corresponding left eigenvectors; see [97] for an application of deflated Krylov subspace methods with oblique projection in the general non-Hermitian case.

4.3.5. Augmented and deflated Krylov subspace methods

In the previous sections, we have described how either augmentation or deflation can be incorporated into Krylov subspace methods of minimum residual norm type. We note that it is possible to combine simultaneously deflation and augmentation into a single Krylov subspace method. In such a setting, the search space of the Krylov subspace method is now decomposed as

$$S_\ell = \mathcal{W} + \mathcal{K}_m(\hat{A}, \hat{r}_0),$$

where $\mathcal{W}$ is the augmentation space of dimension k, $\hat{A}$ refers to the deflated operator and $\hat{r}_0$ to the deflated residual. As an illustration, we briefly review the GCRO (Generalized Conjugate Residual with Orthogonalization) method due to de Sturler [235].

Methods based on augmentation and deflation

Methods based on both augmentation and deflation have been introduced recently; see, e.g., [23, 235, 236, 258]. We focus here on the Generalized Conjugate Residual with inner Orthogonalization (GCRO) [235], which combines augmentation and deflation judiciously, since it will be useful for later developments.

GCRO belongs to the family of inner-outer methods [15, Ch. 12] where the outer iteration is based on the Generalized Conjugate Residual method (GCR), a minimum residual norm Krylov subspace method proposed by Eisenstat, Elman and Schultz [89] while the inner part is based on GMRES. Following the theoretical framework introduced in [87], GCR maintains a correction subspace spanned by $\mathcal{R}(Z_k)$ and an approximation subspace spanned by $\mathcal{R}(V_k)$, where $Z_k, V_k \in \mathbb{C}^{n\times k}$ satisfy the relations

$$\begin{aligned} A\ Z_k &= V_k, \\ V_k^H\ V_k &= I_k. \end{aligned}$$

The optimal solution of the minimization problem $\min \|b - Ax\|_2$ over the subspace $x_0 + \mathcal{R}(Z_k)$ is then found as $x_k = x_0 + Z_k\ V_k^H\ r_0$. Consequently $r_k = b - Ax_k$ satisfies

$$r_k = r_0 - V_k\ V_k^H r_0 = (I_n - V_k\ V_k^H) r_0, \quad r_k \perp \mathcal{R}(V_k).$$

In [235] de Sturler suggested that the inner iteration takes place in a subspace orthogonal to the outer Krylov subspace. In this inner iteration the following projected linear system is considered

$$(I_n - V_k\ V_k^H) Az = (I_n - V_k\ V_k^H) r_k = r_k.$$

The inner iteration is then based on a deflated linear system with $(I_n - V_k\ V_k^H)$ as orthogonal projection. If a minimum residual norm subspace method is used in the inner iteration to solve this projected linear system approximately, the residuals over both the inner and outer subspaces are minimized. Hence, augmentation is applied in the outer iteration and deflation in the inner part of the method. Numerical experiments (see, e.g., [235] and [107, Chapter 1]) indicate that the resulting method may in some cases perform better than other inner-outer methods (without orthogonalization). We mention that the augmentation subspace can be based on spectral approximate invariant subspace information. This leads to the GCRO with deflated restarting method (GCRO-DR) [197] that uses Harmonic Ritz information to define the augmentation subspace as in Section 4.3.2. In [236] de Sturler proposed defining an augmentation subspace based on information other than approximate spectral invariant subspace. At the end of each cycle, the strategy (named GCRO with optimal truncation (GCROT)) decides which part of the current global search subspace to keep to define the new augmentation subspace such that the smallest inner residual norm is obtained. This truncation is done by examining angles between subspaces and requires specification of six different parameters that affect the truncation. We refer to [236] for a complete derivation of the method and numerical experiments (see also [87, Section 4.5]).

4.3.6. Flexible GCRO with deflated restarting: FGCRO-DR

We present next a flexible Krylov subspace method based on *augmentation and deflation*, which consists of our second contribution of this chapter. We introduce both the mathematical derivation of the method and then give an algorithmic description.

Motivations

Formulation

We assume that a generalized Arnoldi relation of type (4.4) holds. As in Section 4.3.3, an important point is to specify what harmonic Ritz information is selected. Given a certain matrix $W_\ell \in \mathbb{C}^{n\times\ell}$ to be specified later on, such as $\mathcal{R}(W_\ell) = \mathcal{R}(V_\ell)$, the deflation procedure relies on the use of k harmonic Ritz vectors $Y_k = W_\ell P_k$ of $AZ_\ell W_\ell^\dagger$ with respect to $\mathcal{R}(W_\ell)$, where $Y_k \in \mathbb{C}^{n\times k}$ and $P_k \in \mathbb{C}^{\ell\times k}$. We point out that the calculation of $W_\ell^\dagger$ is not needed in the practical implementation of the algorithm. In Lemma 4.5, we describe a useful relation satisfied by the harmonic Ritz vectors.

Lemma 4.5. *In flexible GCRO with deflated restarting, the harmonic Ritz vectors are given by $Y_k = W_\ell P_k$ with corresponding harmonic Ritz values θ_k. The matrix $P_k = [p_1, \cdots, p_k] \in \mathbb{C}^{\ell\times k}$ satisfies the following relation*

$$AZ_\ell P_k = [W_\ell P_k, r_0^{(i-1)}] \begin{bmatrix} \operatorname{diag}(\theta_1, \ldots, \theta_k) \\ \beta_{1\times k} \end{bmatrix}, \tag{4.32}$$

where $r_0^{(i-1)} = V_{\ell+1}(c - \bar{H}_\ell y^*)$ and $\beta_{1\times k} = [\beta_1, \ldots, \beta_k] \in \mathbb{C}^{1\times k}$.

Proof. See Lemma 2.3 of [61]. According to Definition 4.2, the harmonic residual vectors $AZ_\ell W_\ell^\dagger W_\ell p_j - \theta_j W_\ell p_j$ and the residual vector $r_0^{(i-1)} = V_{\ell+1}(c - \bar{H}_\ell y^*)$ all belong to a subspace of dimension $\ell+1$ (spanned by the columns of $V_{\ell+1}$) and are orthogonal to the same subspace of dimension ℓ (spanned by the columns of AZ_ℓ subspace of $\mathcal{R}(V_{\ell+1})$), so they must be collinear. Consequently there exist k coefficients noted $\beta_j \in \mathbb{C}$ with $1 \leq j \leq k$ such that

$$\forall j \in \{1, \ldots, k\} \quad AZ_\ell p_j \quad - \quad \theta_j W_\ell p_j = \beta_j r_0^{(i-1)}. \tag{4.33}$$

Setting $\beta_{1\times k} = [\beta_1, \ldots, \beta_k] \in \mathbb{C}^{1\times k}$, the collinearity expression (4.33) can be written in matrix form as

$$AZ_\ell P_k = [W_\ell P_k, r_0^{(i-1)}] \begin{bmatrix} \operatorname{diag}(\theta_1, \ldots, \theta_k) \\ \beta_{1\times k} \end{bmatrix}.$$

□

Due to the generalized Arnoldi relation (4.4), relation (4.32) can be also expressed as

$$V_{\ell+1}\bar{H}_\ell P_k = [W_\ell P_k, r_0^{(i-1)}] \begin{bmatrix} \operatorname{diag}(\theta_1, \ldots, \theta_k) \\ \beta_{1\times k} \end{bmatrix}. \tag{4.34}$$

If required, $\beta_{1\times k}$ can be deduced from (4.34) by

$$(c-\bar{H}_\ell y^*)^H(\bar{H}_\ell P_k - V_{\ell+1}^H W_\ell P_k \operatorname{diag}(\theta_1, \ldots, \theta_k)) = (c-\bar{H}_\ell y^*)^H(c-\bar{H}_\ell y^*)\beta_{1\times k}. \tag{4.35}$$

Next, the QR factorization of the $(\ell+1)\times k$ matrix $\bar{H}_\ell P_k$ appearing in relation (4.34) is performed as $\bar{H}_\ell P_k = QR$ with $Q \in \mathbb{C}^{(\ell+1)\times k}$ and $R \in \mathbb{C}^{k\times k}$.

Proposition 4.6. *In flexible GCRO with deflated restarting, the relation $AZ_k = V_k$ with $V_k^H V_k = I_k$ holds at the i-th cycle with matrices $Z_k, V_k \in \mathbb{C}^{n\times k}$ defined as*

$$\begin{aligned} Z_k &= Z_\ell P_k R^{-1}, \\ V_k &= V_{\ell+1} Q, \end{aligned}$$

where $V_{\ell+1}$ and Z_ℓ refer to matrices obtained at the end of the $(i-1)$-th cycle. In addition $V_k^H r_0^{(i-1)} = 0$ holds during the i-th cycle.

Proof. See Proposition 2 of [61]. By using information related to the QR factorization of $\bar{H}_\ell P_k$ and the generalized Arnoldi relation (4.4) exclusively, we obtain

$$\begin{aligned} A\ Z_k &= AZ_\ell P_k R^{-1}, \\ &= V_{\ell+1}\bar{H}_\ell P_k R^{-1}, \\ &= V_{\ell+1} Q, \\ &= V_k. \end{aligned}$$

Since both $V_{\ell+1}$ and Q have orthonormal columns, V_k satisfies $V_k^H V_k = I_k$. Finally, since $r_0^{(i-1)}$ is the optimum residual at the $i-1$-th cycle, i.e. $(AZ_\ell)^H r_0^{(i-1)} = 0$ we obtain

$$\begin{aligned} P_k^H (AZ_\ell)^H r_0^{(i-1)} &= 0, \\ (V_{\ell+1}\bar{H}_\ell P_k)^H r_0^{(i-1)} &= 0, \\ R^H V_k^H r_0^{(i-1)} &= 0. \end{aligned}$$

This finally shows that $V_k^H r_0^{(i-1)} = 0$ since R is assumed to be nonsingular. □

To complement the subspaces, the inner iteration is based on the approximate solution of

$$(I_n - V_k\ V_k^H)Az = (I_n - V_k\ V_k^H)r_0^{(i-1)} = r_0^{(i-1)},$$

where the last equality is due to Proposition 4.6. For that purpose, FGCRO-DR then carries out m steps of the Arnoldi method with variable preconditioning leading to

$$\begin{aligned} (I_n - V_k V_k^H)\ A\ [z_{k+1}, \cdots, z_{m+k}] &= [v_{k+1}, \cdots, v_{m+k+1}]\ \bar{H}_m \\ (I_n - V_k V_k^H)\ A\ Z_m &= V_{m+1}\ \bar{H}_m \end{aligned}$$

with $v_{k+1} = r_0^{(i-1)}/\|r_0^{(i-1)}\|_2$. At the end of the cycle this gives the generalized Arnoldi relation

$$\begin{aligned} A\ Z_{k+m} &= V_{k+m+1} \begin{bmatrix} I_k & V_k^H\ A\ Z_m \\ 0_{m+1\times k} & \bar{H}_m \end{bmatrix} \\ A\ Z_{m+k} &= V_{m+k+1}\ \bar{H}_{m+k}, \end{aligned}$$

where $Z_{m+k} \in \mathbb{C}^{n\times(m+k)}$, $V_{m+k+1} \in \mathbb{C}^{n\times(m+k+1)}$ and $\bar{H}_{m+k} \in \mathbb{C}^{(m+k+1)\times m}$. At the end of the i-th cycle, an approximate solution $x_0^{(i)} \in \mathbb{C}^n$ is then found by minimizing the residual norm $\|b - A(x_0^{(i-1)} + Z_{m+k}y)\|_2$ over the space $x_0^{(i-1)} + \mathcal{R}(Z_\ell)$, the corresponding residual being $r_0^{(i)} = b - Ax_0^{(i)}$, with $r_0^{(i)} \in \mathcal{R}(V_{\ell+1})$.

Algorithm

Flexible GCRO(m,k) is depicted in Algorithm 4.5. We also refer the reader to [61] for additional comments on the computational cost of FGCRO-DR and a detailed comparison with the flexible variant of GMRES-DR. Numerical results are also provided showing the efficiency of FGRO-DR. When a fixed right preconditioner is used, GMRES-DR and GCRO-DR are equivalent. When variable preconditioning is considered, it is however worthwhile to note that FGMRES-DR and FGCRO-DR are only equivalent if a certain collinearity condition given in [61, Theorem 3.6] is satisfied. Finally, we note that the extension of GCROT to the case of variable preconditioning has been proposed in [146] with application to aerodynamics.

Algorithm 4.5 Flexible GCRO(m,k)

Input: Assume that the following is given:

- $A \in \mathbb{C}^{n\times n}$ ▷ operator
- $b \in \mathbb{C}^n, x_0 \in \mathbb{C}^n$ ▷ right-hand side and initial guess
- $M_j^{-1} \in \mathbb{C}^{n\times n}$ for $j \in \{1,\dots,\ell\}$ ▷ variable preconditioning operators
- $cycle_{\max}$ ▷ maximal number of cycles allowed
- $tol > 0$ ▷ convergence threshold

Settings: Let $r_0 = b - Ax_0$, $\beta = \|r_0\|_2$, $c = [\beta, 0_{1\times\ell}]^T$ where $c \in \mathbb{C}^{\ell+1}$, $v_1 = r_0/\beta$, $i \leftarrow 0$.
Apply FGMRES(ℓ) to obtain $\bar{H}_\ell$, Z_ℓ, $V_{\ell+1}$ such that $AZ_\ell = V_{\ell+1}\bar{H}_\ell$, $y^* = \arg\min_{y\in\mathbb{C}^\ell} \|c - \bar{H}_\ell y\|_2$
$x_0^{(0)} = x_0 + Z_\ell y^*$
$r_0^{(0)} = b - Ax_0^{(0)} = V_{\ell+1}(c - \bar{H}_\ell)y^*), W_\ell = V_\ell$
Define $Y = Z_m R^{-1}$ and $W = V_{m+1}Q$ and select k columns of Y and W into Y_k and W_k with $W_k^H W_k = I_k$ and $W_k^H r_0^{(0)} = 0$
for $cycle = 1, cycle_{\max}$ **do**
    Set $\beta = \|r_0^{(i)}\|_2$ and $v_{k+1} = r_0^{(i)}/\beta$
    for $j = 1,\dots,m$ **do**
        Completion of V_{j+1}, Z_j and H_j: Apply Modified Gram-Schmidt to obtain $V_{j+1} \in \mathbb{C}^{n\times(j+1)}$ $Z_j \in \mathbb{C}^{n\times j}$ and the matrix $\bar{H}_j \in \mathbb{C}^{(j+1)\times j}$ such that:

$$(I_n - W_k W_k^H)AZ_j = V_{j+1}\bar{H}_j \quad \text{with} \quad V_{j+1}^H V_{j+1} = I_{j+1}.$$

        Define $Z_{k+j} = [Y_k, Z_j]$, $V_{k+j+1} = [W_k, V_{j+1}]$ and $\bar{H}_{k+j} = \begin{bmatrix} I_k & W_k^H AZ_j \\ 0_{(j+1)\times k} & \bar{H}_j \end{bmatrix}$
        Define $c \in \mathbb{C}^{j+k+1}$ such that $c = V_{k+j+1}^H r_0^{(i)}$ and $\bar{H}_{k+j} = QR$ with $Q \in \mathbb{C}^{(k+j+1)\times(k+j)}$ and $R \in \mathbb{C}^{(k+j)\times(k+j)}$
        Solve the minimization problem $y^\star = \operatorname{argmin}_{y\in\mathbb{C}^{j+k}} \|c - \bar{H}_{k+j}y\|_2$;
        if $\|c - \bar{H}_{k+j}y^\star\|_2 \le tol$ **then**
            Compute $x_0^{(i+1)} = x_0^{(i)} + Z_{j+k}y^*$; stop;
        end if
    end for
    $x_0^{(i+1)} = x_0^{(i)} + Z_{m+k}y^*$
    $r_0^{(i+1)} = b - Ax_0^{(i+1)}$
    Define $Y = Z_{m+k}R^{-1}$ and $W = V_{m+k+1}Q$ and select k columns of Y and W into Y_k and W_k such that $AY_k = W_k$ with $W_k^H W_k = I_k$ and $W_k^H r_0^{(i+1)} = 0$
    $i \leftarrow i + 1$
end for

4.4. Brief background on block Krylov subspace methods

We briefly describe a few basic properties of block Krylov subspace methods for the solution of (4.3). These notions will be useful for later developments. We refer the reader to the monograph [217] and the survey paper [136] for further details on block Krylov subspace methods.

4.4.1. Problem setting

We consider block Krylov space methods for the solution of linear systems of equations with p right-hand sides provided at the same time, thus:

$$AX = B \tag{4.36}$$

where $A \in \mathbb{C}^{n\times n}$ is assumed to be a nonsingular matrix of large dimension, $B \in \mathbb{C}^{n\times p}$ is full rank and $X \in \mathbb{C}^{n\times p}$. Sparse direct methods based on Gaussian elimination are usually the method of choice when addressing the solution of (4.36) [73, 83]. However, both the complexity of state-of-the-art sparse direct methods in the numerical factorization phase and the related large memory requirements are still considered as the main hurdles for successfully handling linear systems of millions of unknowns. In the sequel, we describe purely iterative methods based on block Krylov space methods that are especially useful when the preconditioning operation is known to be expensive and the dimension of the problem n is large. Finally, although the number of right-hand sides p might be relatively large, we assume here that n is always much larger.

4.4.2. Basic properties of block Krylov subspace methods

In the case of no preconditioning, as stated in [136, 139], a block Krylov subspace method for solving the p systems is an iterative method that generates approximations $X_m \in \mathbb{C}^{n\times p}$ with $m \in \mathbb{N}, (m > 0)$ such that

$$X_m - X_0 \in \mathcal{K}_m^{\square}(A, R_0)$$

where the block Krylov subspace $K_m^{\square}(A, R_0)$ is defined as

$$\mathcal{K}_m^{\square}(A, R_0) = \left\{ \sum_{k=0}^{m-1} A^k R_0 \gamma_k, \ \forall \ \gamma_k \in \mathbb{C}^{p\times p}, \ \text{with } k \mid 0 \le k \le m-1 \right\} \subset \mathbb{C}^{n\times p}.$$

When the right-hand sides are available simultaneously, block Krylov methods are appealing for at least two reasons. Firstly, they enable the systematic use of operations on a block of vectors instead of on a single vector. Depending on the structure of A, this may considerably reduce the number of memory accesses ([22], [164, Section 3.7.2.3]). Secondly, by construction, the block Krylov space $\mathcal{K}_m^{\square}(A, R_0)$ contains all Krylov subspaces generated by each initial residual $\mathcal{K}_m(A, R_0(:, i))$ for i such that $1 \le i \le p$ and all possible linear combinations of the vectors contained in these subspaces. Thus, contrary to the single right-hand side case ($p = 1$), the solution of each linear system is

sought in a potentially richer space, leading hopefully to a reduction in iteration count. We refer the reader to [136] for a recent overview on block Krylov subspace methods and note that most of the standard Krylov subspace methods have a block counterpart (see, e.g., block GMRES [256], block BiCGStab [135] and block QMR [108]).

4.4.3. Block flexible GMRES method

Since variable preconditioning is used, flexible variants of block Krylov subspace methods have to be considered in our setting. As for the single-right hand side case, we have decided to focus only on block flexible Krylov subspace methods based on a norm minimization property. Hence, we briefly describe some basic properties of the flexible variant of block restarted GMRES [91] since it will be the basis for later developments.

Formulation

We introduce a flexible variant relying on a block version of the Arnoldi method. The orthogonalization scheme chosen is block modified Gram-Schmidt, although it is clear that one can change it at will with similar convergence effects as for the GMRES algorithm in floating-point arithmetic. The block orthogonalization procedure used in the flexible setting, where M_j^{-1} denotes the preconditioning operator at step j $(1 \leq j \leq m)$ is based on the following important relation.

Definition 4.7. Generalized block Arnoldi relation. The flexible block Arnoldi method given in Algorithm 4.6 (Section 4.4.3) leads to the following relation (later called generalized block Arnoldi relation), for $1 \leq j \leq m$,

$$A\,[Z_1,\ldots,Z_j] \;=\; [V_1,V_2,\ldots,V_{j+1}] \begin{bmatrix} H_{1,1} & H_{1,2} & \ldots & H_{1,j} \\ H_{2,1} & H_{2,2} & \ldots & H_{2,j} \\ 0_{p\times p} & H_{3,2} & \ldots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & H_{j+1,j} \end{bmatrix}.$$

Equivalently with the notation introduced in Algorithm 4.6 line 10, the orthogonalization procedure produces matrices $\mathcal{Z}_j \in \mathbb{C}^{n\times jp}$, $\mathcal{V}_{j+1} \in \mathbb{C}^{n\times(j+1)p}$ and $\bar{\mathcal{H}}_j \in \mathbb{C}^{(j+1)p\times jp}$ which satisfy

$$A\mathcal{Z}_j = \mathcal{V}_{j+1}\,\bar{\mathcal{H}}_j. \tag{4.37}$$

It should be noticed that $\bar{\mathcal{H}}_j$ is no longer a Hessenberg matrix but a block Hessenberg matrix. More precisely, its block sub-diagonal consists of upper triangular blocks of size $p \times p$. BFGMRES(m) (given in Algorithm 4.7) uses the flexible block version of the Arnoldi method with modified block Gram-Schmidt presented in Algorithm 4.6. In Algorithm 4.7, we denote by $\mathcal{B}_j \in \mathbb{C}^{(j+1)p\times p}$ the representation of the block residual

$R_0 = B - AX_0$ in the $\mathcal{V}_{j+1}$ basis ($R_0 = \mathcal{V}_{j+1}\mathcal{B}_j$) and by $Y_j \in \mathbb{C}^{jp\times p}$ the solution of the following minimization problem in the Frobenius norm

$$\mathcal{P}_r : Y_j = \text{argmin}_{Y\in\mathbb{C}^{jp\times p}} \|\mathcal{B}_j - \bar{\mathcal{H}}_j Y\|_F. \tag{4.38}$$

Finally, we recall a convergence property of block GMRES shown in [256] extended to the case of block flexible GMRES. Proposition 4.8 shows that the block flexible GMRES method minimizes the Euclidean norm of the residual of each linear system.

Proposition 4.8. *In block flexible GMRES (BFGMRES(m), Algorithm 4.7) solving the reduced minimization problem $\mathcal{P}_r$ (4.38) amounts to minimizing the Frobenius norm of the block true residual $\|B-AX\|_F$ over the space $X_0+\mathcal{R}(\mathcal{Z}_jY)$ at iteration j $(1 \le j \le m)$ of a given cycle, i.e.*

$$\begin{aligned}
&\text{argmin}_{Y\in\mathbb{C}^{jp\times p}} \|\mathcal{B}_j - \bar{\mathcal{H}}_j Y\|_F = \text{argmin}_{Y\in\mathbb{C}^{jp\times p}} \|B - A(X_0 + \mathcal{Z}_j Y)\|_F,\\
&\min_{Y\in\mathbb{C}^{jp\times p}} \|\mathcal{B}_j - \bar{\mathcal{H}}_j Y\|_F = \min_{Y\in\mathbb{C}^{jp\times p}} \|B - A(X_0 + \mathcal{Z}_j Y)\|_F.
\end{aligned} \tag{4.39}$$

Furthermore, solving the reduced minimization problem $\mathcal{P}_r$ (4.38) is also equivalent to minimizing the Euclidean norm of each linear system over the space $X_0(:,l) + \mathcal{R}(\mathcal{Z}_j)$ $(1 \le l \le p)$ at iteration j $(1 \le j \le m)$.

Proof. See Proposition 1 of [59]. □

Algorithms

Block Arnoldi Algorithm 4.6 introduces the block Arnoldi procedure with block modified Gram-Schmidt. Algorithm 4.6 proceeds by orthonormalizing AZ_j against all the previous preconditioned Krylov directions. The block modified Gram-Schmidt version is presented in Algorithm 4.6, but a version of block Arnoldi due to Ruhe [213] or block Householder orthonormalization [21, 217, 237] could be used as well.

Algorithm 4.6 Flexible block Arnoldi with block Modified Gram-Schmidt: computation of $\mathcal{V}_{j+1}$, $\mathcal{Z}_j$ and $\bar{\mathcal{H}}_j$ for $1 \leq j \leq m$ with $V_1 \in \mathbb{C}^{n\times p}$ such that $V_1^H V_1 = I_p$

Input: Assume that the following is given:

- $A \in \mathbb{C}^{n\times n}$ ▷ operator
- $V_1 \in \mathbb{C}^{n\times p}$ ▷ block of vectors such that $V_1^H V_1 = I_p$
- $M_j^{-1} \in \mathbb{C}^{n\times n}$ for $j \in \{1, \ldots, m\}$ ▷ variable preconditioning operators

for $j = 1, \ldots, m$ **do**
  $Z_j = M_j^{-1} V_j$
  $S = A Z_j$
  for $i = 1, \ldots, j$ **do**
    $H_{i,j} = V_i^H S$
    $S = S - V_i H_{i,j}$
  end for
  Compute the QR decomposition of S as $S = QR$ with $Q \in \mathbb{C}^{n\times p}$ and $R \in \mathbb{C}^{p\times p}$
  Set $V_{j+1} = Q$, $H_{j+1,j} = R$ and $H_{i,j} = 0_{p\times p}$ for $i > j+1$
  Define $\mathcal{Z}_j = [Z_1, \ldots, Z_j]$, $\mathcal{V}_{j+1} = [V_1, \ldots, V_{j+1}]$, $\bar{\mathcal{H}}_j = (H_{k,l})_{1\leq k\leq j+1, 1\leq l\leq j}$
end for

BFGMRES(m) Algorithm 4.7 introduces the block flexible GMRES method. This algorithm is named BFGMRES(m), where m denotes the maximum number of iterations performed in a given cycle.

Computational cost of a cycle

We summarize in Table 4.1 the costs incurred during a given cycle of BFGMRES(m) (considering Algorithms 4.6 and 4.7), excluding the cost of the m matrix-vector products and m preconditioning operations which are problem dependent. We have included the costs proportional to both the size of the original problem n and the number of right-hand sides p, assuming a QR factorization based on modified Gram-Schmidt and a Golub-Reinsch SVD[1]; see, e.g, [127, Section 5.4.5] and [147, Appendix C] for further details on operation counts. The total cost of a given cycle is then found to grow as $C_1 np^2 + C_2 np$ as shown in Table 4.1.

4.5. Block flexible Krylov subspace methods including block size reduction at restart

Motivations Although potentially appealing as discussed in Section 4.4.2, block (flexible) GMRES based algorithms are known to be computationally expensive due to the cost of orthogonalization which behaves as $2np^2m^2$[136]. Thus a primary concern when deriving those variants is to remove useless information for the convergence as soon

[1]The Golub-Reinsch SVD decomposition $R = U\Sigma V^H$ with $R \in \mathbb{C}^{m\times n}$ requires $4mn^2 + 8n^3$ operations when only Σ and V have to be computed.

Algorithm 4.7 Block Flexible GMRES (BFGMRES(m))

Input: Assume that the following is given:

- $A \in \mathbb{C}^{n\times n}$ ▷ operator
- $B, X_0 \in \mathbb{C}^{n\times p}$ ▷ right-hand sides and initial guess
- $M_j^{-1} \in \mathbb{C}^{n\times n}$ for $j \in \{1,\dots,m\}$ ▷ variable preconditioning operators
- $cycle_{\max}$ ▷ maximal number of cycles allowed
- $tol > 0$ ▷ convergence threshold

1: Compute the initial block residual $R_0 \in \mathbb{C}^{n\times p}$ as $R_0 = B - AX_0$
2: **for** $cycle = 1,\dots,cycle_{\max}$ **do**
3: Compute the QR decomposition of R_0 as $R_0 = QT$ with $Q \in \mathbb{C}^{n\times p}$ and $T \in \mathbb{C}^{p\times p}$
4: Set $V_1 = Q$ and $\mathcal{B}_k = \begin{bmatrix} T \\ 0_{kp\times p} \end{bmatrix}$, $1 \le k \le m$.
5: **for** $j = 1,\dots,m$ **do**
6: *Completion of $\mathcal{V}_{j+1}$, $\mathcal{Z}_j$ and $\bar{\mathcal{H}}_j$:* Apply Algorithm 4.6 from line 2 to 10 with variable preconditioning ($Z_j = M_j^{-1}V_j$, $1 \le j \le m$) to obtain $\mathcal{V}_{j+1} \in \mathbb{C}^{n\times(j+1)p}$, $\mathcal{Z}_j \in \mathbb{C}^{n\times jp}$ and the matrix $\bar{\mathcal{H}}_j \in \mathbb{C}^{(j+1)p\times jp}$ such that

$$A\mathcal{Z}_j = \mathcal{V}_{j+1}\bar{\mathcal{H}}_j \quad \text{with} \quad \mathcal{V}_{j+1}^H\mathcal{V}_{j+1} = I_{(j+1)p}.$$

7: Solve the minimization problem $Y_j = \operatorname{argmin}_{Y\in\mathbb{C}^{jp\times p}} \|\mathcal{B}_j - \bar{\mathcal{H}}_j Y\|_F$
8: **if** $\|\mathcal{B}_j(:,l) - \bar{\mathcal{H}}_j Y_j(:,l)\|_2 / \|B(:,l)\|_2 \le tol,\ \forall\ l \mid 1 \le l \le p$ **then**
9: Compute $X_j = X_0 + \mathcal{Z}_j Y_j$; stop
10: **end if**
11: **end for**
12: Compute $X_m = X_0 + \mathcal{Z}_m Y_m$ and $R_m = B - AX_m$
13: Set $R_0 = R_m$ and $X_0 = X_m$
14: **end for**

Step	Computational cost of a cycle
Computation of R_0	np
QR factorization of R_0	$2np^2 + 5np$
Block Arnoldi procedure	$2nm(m+2)p^2 + (5mn + \frac{m(m+1)}{2}n)p$
Computation of X_m	$np + nmp^2$
Total	$np^2[2m^2 + 5m + 2] +$ $np[5m + 7 + (m+1)\frac{m}{2}]$

Table 4.1.: Computational cost of a cycle of BFGMRES(m). This excludes the cost of matrix-vector operations and preconditioning operations. Table 3.1 of [59].

as possible during the iterative procedure. This assumes the inclusion of strategies for detecting when a linear combination of the p systems has approximately converged.

The first obvious strategy for removing unneeded information from a block Krylov subspace is called initial deflation in [136]. It consists in detecting linear dependency in the block right-hand side B or in the initial block residual R_0 ([136, Section 12] and [164, Section 3.7.2]). This requires us to compute numerical ranks using rank-revealing QR-factorizations [52] or singular value decompositions (SVD) [127] according to a certain deflation tolerance [148]. In Section 4.5, we examine in detail a strategy related to block size reduction. It aims at removing unneeded information *at each restart* of the block flexible GMRES method.

4.5.1. Formulation

The block flexible restarted GMRES with block size reduction later named BFGMRES-D(m) is presented in Algorithm 4.8 given in Section 4.5.2. Hereafter, we outline how approximate block size reduction has been introduced and thus describe a given cycle of the method (lines 4 to 19 in Algorithm 4.8). The block size reduction procedure detects approximate linear dependency in the block true residual. For that purpose, given a QR-factorization of the scaled block true residual $R_0\,D^{-1} = QT$ where $D \in \mathbb{C}^{p\times p}$ is defined as $D = diag(d_1, \ldots, d_p)$ with $d_l = \|B(:,l)\|_2$ $(1 \le l \le p)$, a singular value decomposition of the upper triangular matrix $T \in \mathbb{C}^{p\times p}$ is performed which leads to the following relation

$$T = U\Sigma W^H \tag{4.40}$$

where $U \in \mathbb{C}^{p\times p}$, $W \in \mathbb{C}^{p\times p}$ are unitary and $\Sigma \in \mathbb{C}^{p\times p}$ is diagonal. The use of diagonal scaling with matrix D enables the convergence detection of the true block residual scaled by the norm of the right-hand sides. Block size reduction consists in selecting relevant information from the decomposition (4.40). Indeed, we determine a subset of the singular values of T according to the following condition

$$\sigma_l(T) > \varepsilon_d\ tol \quad \forall\ l \text{ such that } 1 \le l \le p_d \tag{4.41}$$

where ε_d is a real positive parameter less than or equal to one. This leads to the following decomposition of the diagonal matrix Σ

$$\Sigma = \begin{bmatrix} \Sigma_+ & 0_{p_d\times(p-p_d)} \\ 0_{(p-p_d)\times p_d} & \Sigma_- \end{bmatrix}$$

with $\Sigma_+ \in \mathbb{C}^{p_d\times p_d}$ defined as $\Sigma_+ = \Sigma(1:p_d, 1:p_d)$ and $\Sigma_- \in \mathbb{C}^{(p-p_d)\times(p-p_d)}$ as $\Sigma_- = \Sigma(p_d+1:p, p_d+1:p)$. Due to the approximate block size reduction condition (4.41), we note that

$$\|\Sigma_+\|_2 > \varepsilon_d\ tol \quad \text{and} \quad \|\Sigma_-\|_2 \le \varepsilon_d\ tol.$$

Furthermore the scaled block true residual $R_0\,D^{-1}$ can be written as

$$\begin{aligned} R_0\,D^{-1} &= Q\ [U_+\ U_-] \begin{bmatrix} \Sigma_+ & 0_{p_d\times(p-p_d)} \\ 0_{(p-p_d)\times p_d} & \Sigma_- \end{bmatrix} [W_+\ W_-]^H, \\ R_0\,D^{-1} &= Q\ U_+\Sigma_+W_+^H + Q\ U_-\Sigma_-W_-^H \end{aligned} \tag{4.42}$$

where we set $U_+ \in \mathbb{C}^{p\times p_d}$ as $U_+ = U(:, 1 : p_d)$ and $W_+ \in \mathbb{C}^{p\times p_d}$ as $W_+ = W(:, 1 : p_d)$. Similarly, we define $U_- \in \mathbb{C}^{p\times(p-p_d)}$ as $U_- = U(:, p_d + 1 : p)$ and $W_- \in \mathbb{C}^{p\times(p-p_d)}$ as $W_- = W(:, p_d + 1 : p)$. U_+, W_+ and Σ_+ denote the quantities effectively considered in a given cycle of Algorithm 4.8, while U_-, W_- and Σ_- are put aside due to block size reduction. Indeed since $W = [W_+, W_-]$ is unitary, it is straightforward to see from (4.42) that

$$\|R_0\, D^{-1} W_-\|_2 \leq \varepsilon_d\ tol.$$

If block size reduction is active in this cycle ($p_d < p$), *only* p_d linear systems will be considered which may yield a significant reduction in terms of operations. Given $V_1 = QU_+$ the generalized block Arnoldi method with block Modified Gram-Schmidt (Algorithm 4.6) is applied to obtain $\mathcal{Z}_j \in \mathbb{C}^{n\times jp_d}$, $\mathcal{V}_{j+1} \in \mathbb{C}^{n\times(j+1)p_d}$ and $\bar{\mathcal{H}}_j \in \mathbb{C}^{(j+1)p_d\times jp_d}$ which satisfy

$$A\mathcal{Z}_j = \mathcal{V}_{j+1}\,\bar{\mathcal{H}}_j. \tag{4.43}$$

We denote by $\mathcal{B}_j \in \mathbb{C}^{(j+1)p_d\times p_d}$ the representation of the scaled block residual in the $\mathcal{V}_{j+1}$ basis ($\mathcal{V}_{j+1}\mathcal{B}_j = QU_+$) and by $Y_j \in \mathbb{C}^{jp_d\times p_d}$ the solution of the reduced minimization problem

$$\mathcal{P}_r^d: \quad Y_j = \mathrm{argmin}_{Y\in\mathbb{C}^{jp_d\times p_d}} \|\mathcal{B}_j - \bar{\mathcal{H}}_j Y\|_F. \tag{4.44}$$

Proposition 4.9. *In block flexible GMRES with block size reduction (BFGMRES-D(m), Algorithm 4.8) solving the reduced minimization problem $\mathcal{P}_r^d$ (4.44) amounts to minimizing the Frobenius norm of the block true residual $\|B - AX\|_F$ over the space $X_0 + \mathcal{R}(\mathcal{Z}_j Y \Sigma_+ W_+^H D)$ at iteration j $(1 \leq j \leq m)$ of a given cycle, i.e.,*

$$\begin{aligned}\mathrm{argmin}_{Y\in\mathbb{C}^{jp_d\times p_d}} \|\mathcal{B}_j - \bar{\mathcal{H}}_j Y\|_F &= \mathrm{argmin}_{Y\in\mathbb{C}^{jp_d\times p_d}} \|B - A(X_0 + \mathcal{Z}_j Y \Sigma_+ W_+^H D)\|_F,\\ &= \mathrm{argmin}_{Y\in\mathbb{C}^{jp_d\times p_d}} \|R_0 D^{-1} - A\mathcal{Z}_j Y \Sigma_+ W_+^H\|_F. \end{aligned} \tag{4.45}$$

Proof. See Proposition 2 of [59]. □

Due to Proposition 4.9, the approximate solution that is based on a generalized minimum Frobenius norm approach is obtained as

$$X_j = X_0 + \mathcal{Z}_j Y_j \Sigma_+ W_+^H D$$

at the end of the cycle ($j = m$) or before if the stopping criterion is satisfied at iteration j. Proposition 4.9 also implies the nonincreasing behaviour of the Frobenius norm of the block residual in BFGMRES-D(m).

4.5.2. Algorithms

BFGMRES-D(m) Algorithm 4.8 introduces the block flexible GMRES method with deflation at restart. This algorithm is later named BFGMRES-D(m) where m denotes the maximum number of iterations performed in a given cycle. The suffix "D" is used to emphasise that the method is based on deflation (i.e. block size reduction) at each restart, as described above.

Algorithm 4.8 Block Flexible GMRES with SVD based deflation (BFGMRES-D(m))

Input: Assume that the following is given:

- $A \in \mathbb{C}^{n\times n}$ ▷ operator
- $B, X_0 \in \mathbb{C}^{n\times p}$ ▷ right-hand side and initial guess
- $M_j^{-1} \in \mathbb{C}^{n\times n}$ for $j \in \{1,\ldots,m\}$ ▷ variable preconditioning operators
- $cycle_{\max}$ ▷ maximal number of cycles allowed
- $tol > 0$ ▷ convergence threshold
- $\varepsilon_d > 0$ ▷ deflation threshold
- ε_q ▷ quality of convergence threshold

1: Define the diagonal matrix $D \in \mathbb{C}^{p\times p}$ as $D = diag(d_1,\ldots,d_p)$ with $d_l = \|B(:,l)\|_2$ for l such that $1 \le l \le p$
2: Compute the initial block residual $R_0 = B - AX_0$
3: **for** $cycle = 1,\ldots,cycle_{\max}$ **do**
4: Compute the QR decomposition of R_0D^{-1} as $R_0D^{-1} = QT$ with $Q \in \mathbb{C}^{n\times p}$ and $T \in \mathbb{C}^{p\times p}$
5: Compute the SVD of T as $T = U\ \Sigma\ W^H$
6: Select p_d singular values of T such that $\sigma_l(T) > \varepsilon_d\ tol$ for all l such that $1 \le l \le p_d$
7: Define $V_1 \in \mathbb{C}^{n\times p_d}$ as $V_1 = QU(:,1:p_d)$
8: Let $\mathcal{B}_k = \begin{bmatrix} I_{p_d} \\ 0_{kp_d\times p_d} \end{bmatrix}$, $1 \le k \le m$
9: **for** $j = 1,\ldots,m$ **do**
10: *Completion of $\mathcal{V}_{j+1}$, $\mathcal{Z}_j$ and $\bar{\mathcal{H}}_j$ (see Algorithm 4.6):* Apply Algorithm 4.6 from line 2 to 10 with variable preconditioning ($Z_j = M_j^{-1}V_j$, $1 \le j \le m$) to obtain $\mathcal{V}_{j+1} \in \mathbb{C}^{n\times(j+1)p_d}$, $\mathcal{Z}_j \in \mathbb{C}^{n\times jp_d}$ and the matrix $\bar{\mathcal{H}}_j \in \mathbb{C}^{(j+1)p_d\times jp_d}$ such that:

$$A\mathcal{Z}_j = \mathcal{V}_{j+1}\bar{\mathcal{H}}_j \quad \text{with} \quad \mathcal{V}_{j+1}^H\mathcal{V}_{j+1} = I_{(j+1)p_d}.$$

11: Solve the minimization problem $Y_j = \operatorname{argmin}_{Y\in\mathbb{C}^{jp_d\times p_d}} \|\mathcal{B}_j - \bar{\mathcal{H}}_jY\|_F$;
12: Compute $\mathcal{R}_j = (\mathcal{B}_j - \bar{\mathcal{H}}_jY_j)\Sigma(1:p_d,1:p_d)W(1:p,1:p_d)^H$
13: **if** $\|\mathcal{R}_j(:,l)\|_2 \le \varepsilon_q\ tol,\ \forall\ l \mid 1 \le l \le p$ **then**
14: Compute $X_j = X_0 + \mathcal{Z}_jY_j\Sigma(1:p_d,1:p_d)W(1:p,1:p_d)^HD$; stop;
15: **end if**
16: **end for**
17: $X_m = X_0 + \mathcal{Z}_mY_m\Sigma(1:p_d,1:p_d)W(1:p,1:p_d)^H\ D$
18: $R_m = B - AX_m$
19: Set $R_0 = R_m$ and $X_0 = X_m$
20: **end for**

4.5.3. Computational cost of a cycle

We summarize in Table 4.2 the main computational costs occurring during a given cycle of BFGMRES-D(m) (Algorithm 4.8). We have only included the costs proportional to the size of the original problem n which is assumed to be much greater than m and

p in practical applications. This also excludes the costs related to both matrix-vector products and preconditioning operations. The total cost is cubic in p_d (the maximum column size of the block vectors in a given cycle) and linear in n (the dimension of the problem).

Step	Computational cost of a cycle
Computation of $R_0 D^{-1}$	np
QR factorization of $R_0 D^{-1}$	$2np_b^2 + 5np_b$
Computation of V_1	$2nppb$
Computation of T	$14p_b^3$
Block Arnoldi procedure[1]	$2nm(m+2)p_b^2 + (5mn + \frac{m(m+1)}{2}n)p_b$
Computation of X_m	$np + nmpp_b$
Total	$np_b^2[2(m+1)^2]+$ $np_b[p(m+2) + (m+1)\frac{(10+m)}{2}]+$ $n(p+1)$

Table 4.2.: Maximum computational cost of a cycle of BFGMRES-D(m) with $p_b = \min(p, p_d)$. This excludes the cost of matrix-vector operations and preconditioning operations. Table 3.1 of [59].

In terms of maximum memory requirements (proportional to n), BFGMRES-D(m) requires the storage of R_m, X_0, X_m, $\mathcal{V}_{m+1}$ and $\mathcal{Z}_m$ respectively, i.e, $n(2m+1)p + 3np$. This is similar to the maximum storage of BFGMRES(m).

4.5.4. Numerical illustration

We investigate the numerical behaviour of block flexible Krylov subspace methods including deflation at each restart or at each iteration (see [57] and Appendix B.5) on a challenging application in geophysics where the multiple right-hand side situation frequently occurs (full waveform inversion). The source terms correspond to Dirac sources in this example. Thus the block right-hand side $B \in \mathbb{C}^{n \times p}$ is extremely sparse (only one nonzero entry per column) and the initial block residual corresponds to a full rank matrix. We compare BFGMRES(m), BFGMRES-D(m) and BFGMRES-S(m) (see Algorithm 3 of [57]) with a zero initial guess (X_0) and a moderate value of the restart parameter m. The iterative procedures are stopped when the condition

$$\frac{\|B(:,l) - AX(:,l)\|_2}{\|B(:,l)\|_2} \leq tol, \ \forall \ l = 1, \ldots, p,$$

is satisfied. As a preconditioner, we consider the basic two-grid method described in Section 2.5. Due to the approximate solution on the coarse grid, a given cycle is thus expensive. The assumption related to the preconditioner is thus satisfied. In the sequel, we use exactly the same parameters as in Section 2.8 related to the single right-hand side situation; they were described in Section 2.8.1.

Table 4.3 includes, in addition to iterations (It)[2] and preconditioner applications on a single vector $(Prec)$[3], the computational times in seconds (T). Among the different strategies BFGMRES-S(5) always delivers the minimal number of preconditioner applications and computational times (see italic and bold values, respectively, in Table 4.3). This clearly highlights the interest of performing deflation at each iteration both in terms of preconditioner applications and computational operations for this given application. The improvement over BFGMRES(5) ranges from 27% for $p = 4$ to 57% for $p = 128$ which is a very satisfactory behaviour. BFGMRES-S(5) is also found to be competitive with respect to methods incorporating deflation at restart only (a gain of up to 15% in terms of computational time is obtained for instance for $p = 8$). This is a satisfactory improvement since methods including deflation at restart only are already quite efficient in this application as shown in [59].

Acoustic full waveform inversion - $Grid : 433 \times 433 \times 126$									
	$p = 4$			$p = 8$			$p = 16$		
Method	It	Prec	T	It	Prec	T	It	Prec	T
BFGMRES(5)	14	56	622	14	112	631	14	224	668
BFGMRES-D(5)	14	43	489	15	70	401	15	120	371
BFGMRES-S(5)	16	*39*	**452**	16	*57*	**339**	18	*102*	**328**
	$p = 32$			$p = 64$			$p = 128$		
Method	It	Prec	T	It	Prec	T	It	Prec	T
BFGMRES(5)	14	448	713	18	1152	962	19	2432	1187
BFGMRES-D(5)	15	225	371	20	490	422	25	1015	509
BFGMRES-S(5)	19	*181*	**316**	25	*413*	**375**	28	*915*	**497**

Table 4.3.: Acoustic full waveform inversion (SEG/EAGE Overthrust model) at $f = 3.64\ Hz$, with $p = 4$ to $p = 128$ right-hand sides given at once. It denotes the number of iterations, $Prec$ the number of preconditioner applications on a single vector and T denotes the total computational time in seconds. The number of cores is set to $8p$. Table 5.1 of [57].

Figure 4.1 shows the evolution of k_j (number of Krylov directions at iteration j) with convergence for the various block subspace methods in the case of $p = 32$. Regarding BFGMRES-D(5), deflation is performed only at the beginning of each cycle, thus k_j is found to be constant in a given cycle. Variations at each iteration can only happen in BFGMRES-S(5). As expected, k_j monotonically decreases as algorithm converges (see Proposition 3.3 in [57]).

[2] A complete cycle of BFGMRES(m), BFGMRES-R(m) or BFGMRES-S(m) always corresponds to m iterations.

[3] A complete cycle of BFGMRES(m) corresponds to mp preconditioner applications, whereas a complete cycle of either BFGMRES-R(m) or BFGMRES-S(m) corresponds to $\sum_{j=1}^{m} k_{j,c}$ preconditioner applications.

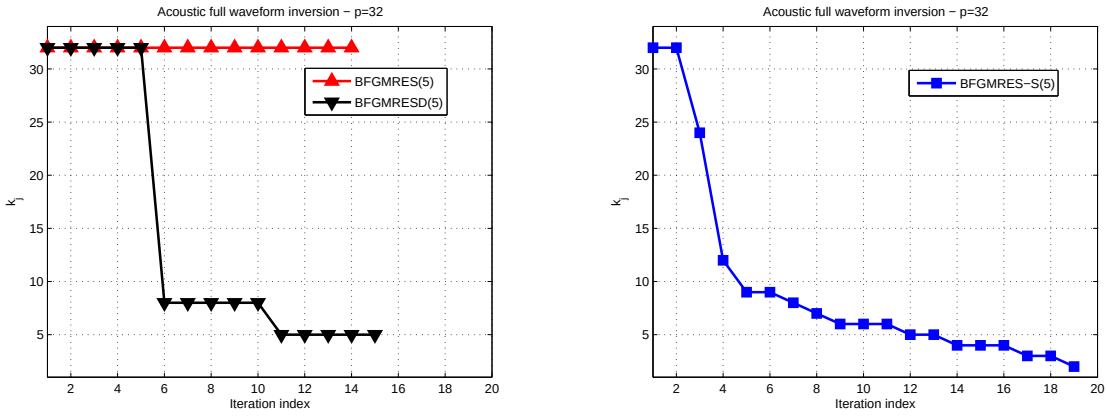


Figure 4.1.: Acoustic full waveform inversion (SEG/EAGE Overthrust model) with $p = 32$. Evolution of k_j (number of Krylov directions at iteration j) versus iterations for $p = 32$ in BFGMRES(5), BFGMRES-D(5) (left part) and BFGMRES-S(5) (right part). Figure 5.1 of [57].

4.6. Additional comments and conclusions

Summary In this chapter, we have focused on certain minimum residual norm Krylov subspace methods for the solution of large linear systems of equations with variable preconditioners. Krylov subspace methods based on augmentation and/or on deflation have been proposed to possibly improve their convergence rate when variable preconditioning is used. We have also considered the case of linear systems with multiple right-hand sides given at the same time and have described in detail advanced block Krylov subspace methods for such a purpose. A realistic application has been provided to illustrate both the performance and benefits of the new numerical methods. We refer the reader to [57, 59, 60, 61, 124] for further comments on numerical experiments.

Collaboration These research projects have been developed with Henri Calandra (TOTAL), Serge Gratton, Xavier Pinel and Rafael Lago. This research has also benefited from interactions or collaborations with Luiz Mariano Carvalho, Luc Giraud, Martin Gutknecht and Julien Langou.

Software realization The Krylov subspace methods that have been proposed in this chapter are generic methods. Implementations in Matlab, and Fortran 90 with BLAS and LAPACK libraries have been realised and integrated into different application codes.

Short-term prospectives As short-term prospectives, we anticipate that a combination of spectral deflation and/or of augmentation with block flexible Krylov subspace methods including block size reduction (at each iteration or at each restart) could be derived as well. As an illustration, we give in Algorithm 4.9 the extension of the flexible

GCRO Krylov subspace method presented in Section 4.3.6 to the multiple right-hand side setting.

The analysis of deflation based on oblique projections could be foreseen in this framework, especially when non-Hermitian matrices are considered. Finally, the solution of linear systems with right-hand sides given in sequence is also a relevant topic that must be considered. Krylov subspace methods based on recycling [119, 197] have been proposed in such a setting; the case of variable preconditioning in the sequence must be studied; see Section 5.4.1 for further details.

Algorithm 4.9 Block Flexible GCRO (BFGCRO(m))

Input: Assume that the following is given:

- $A \in \mathbb{C}^{n\times n}$ ▷ operator
- $B, X_0 \in \mathbb{C}^{n\times p}$ ▷ right-hand side and initial guess
- $M_j^{-1} \in \mathbb{C}^{n\times n}$ for $j \in \{1,\ldots,m\}$ ▷ variable preconditioning operators
- $cycle_{\max}$ ▷ maximal number of cycles allowed
- $tol > 0$ ▷ convergence threshold

1: Define the diagonal matrix $D \in \mathbb{C}^{p\times p}$ as $D = diag(d_1,\ldots,d_p)$ with $d_l = \|B(:,l)\|_2$ for l such that $1 \le l \le p$
2: Compute the initial block residual $R_0 = B - AX_0$
3: Set $\mathcal{Y}_k = []$, $\mathcal{W}_k = []$
4: **for** $cycle = 1,\ldots,cycle_{\max}$ **do**
5: Compute the QR decomposition of R_0D^{-1} as $R_0D^{-1} = QT$ with $Q \in \mathbb{C}^{n\times p}$ and $T \in \mathbb{C}^{p\times p}$
6: Define $V_1 \in \mathbb{C}^{n\times p}$ as $V_1 = Q$
7: **for** $j = 1,\ldots,m$ **do**
8: *Completion of $\mathcal{V}_{j+1}$, $\mathcal{Z}_j$ and $\bar{\mathcal{H}}_j$ (see Algorithm 4.6):* Apply Algorithm 4.6 from line 2 to 10 with variable preconditioning ($Z_j = M_j^{-1}V_j$, $1 \le j \le m$) to obtain $\mathcal{V}_{j+1} \in \mathbb{C}^{n\times(j+1)p}$, $\mathcal{Z}_j \in \mathbb{C}^{n\times jp}$ and the matrix $\bar{\mathcal{H}}_j \in \mathbb{C}^{(j+1)p\times jp}$ such that:

$$(I_n - \mathcal{W}_k\mathcal{W}_k^H)A\mathcal{Z}_j = \mathcal{V}_{j+1}\bar{\mathcal{H}}_j \quad \text{with} \quad \mathcal{V}_{j+1}^H\mathcal{V}_{j+1} = I_{(j+1)p}.$$

9: Define $\mathcal{Z}_{k+j} = [\mathcal{Y}_k, \mathcal{Z}_j]$, $\mathcal{V}_{k+j+1} = [\mathcal{W}_k, \mathcal{V}_{j+1}]$ and $\bar{\mathcal{H}}_{k+j} = \begin{bmatrix} I_{kp} & \mathcal{W}_k^H A\mathcal{Z}_j \\ 0_{(j+1)p\times kp} & \bar{\mathcal{H}}_j \end{bmatrix}$
10: Define $\mathcal{B}_{k+j} \in \mathbb{C}^{(j+k+1)p\times p}$ such that $\mathcal{B}_{k+j} = \begin{bmatrix} 0_{kp\times p} \\ T \\ 0_{jp\times p} \end{bmatrix}$
11: Solve the minimization problem $Y_{k+j} = \operatorname{argmin}_{Y\in\mathbb{C}^{(j+k)p\times p}} \|\mathcal{B}_{k+j} - \bar{\mathcal{H}}_{k+j}Y\|_F$;
12: Compute $\mathcal{R}_{k+j} = \mathcal{B}_{k+j} - \bar{\mathcal{H}}_{k+j}Y_{k+j}$
13: **if** $\|\mathcal{R}_{k+j}(:,l)\|_2 \le tol,\ \forall\ l \mid 1 \le l \le p$ **then**
14: Compute $X_j = X_0 + \mathcal{Z}_{k+j}Y_{k+j}D$; stop;
15: **end if**
16: **end for**
17: $X_m = X_0 + \mathcal{Z}_{k+m}Y_{k+m}D$
18: $R_m = B - AX_m$
19: Recycling: select $\mathcal{Y}_k \in \mathbb{C}^{n\times kp}$, $\mathcal{W}_k \in \mathbb{C}^{n\times kp}$ such that $A\mathcal{Y}_k = \mathcal{W}_k$ with $\mathcal{W}_k^H\mathcal{W}_k = I_{kp}$ and $\mathcal{W}_k^H R_m = 0_{kp\times p}$
20: Set $R_0 = R_m$ and $X_0 = X_m$
21: **end for**

5. Prospectives

5.1. Objective

In this chapter, we present current and future short-term prospectives concerning our main goals defined in the Introduction (see Chapter 1). The forthcoming parallel computers will offer new opportunities to tackle mathematical problems related to the simulation, optimization, control or uncertainty quantification of physical phenomena based on deterministic or stochastic partial differential equations. Numerically stable algorithms must be then adapted, improved or completely redesigned to exploit as much as possible, e.g., the extreme core count of future computing resources. The main objective in this chapter is to briefly present selected research activities in this direction.

5.2. Synopsis

We first address the solution of algebraic linear systems of equations in Section 5.3, where advanced scalable solvers based on multilevel methods are presented. Then, in Section 5.4, we describe relevant strategies for the numerical solution of nonlinear, time-dependent partial differential equations that can either further improve the rate of convergence of multilevel methods or increase the degree of parallelism. We give motivations, describe the problem setting, and specify the currently available results for each topic. Conclusions are briefly drawn in Section 5.5.

5.3. Towards extremely scalable linear solvers

The next two research prospectives that we present are related to the solution of algebraic linear systems of equations. A specific focus on multilevel methods and Krylov subspace methods is proposed. When available, numerical results are presented.

5.3.1. Algebraic multigrid method

Motivations

We consider the class of algebraic multigrid methods (AMG) for the solution of elliptic partial differential equations with possibly variable coefficients. AMG methods were first proposed in [42], [212]; see [247, Appendix A] for a comprehensive survey. They have been shown to deliver *scalable*, sometimes *optimal*, solution methods in such a

setting; see, e.g., [191, 192]. In AMG (Algorithm 5.1), the multigrid hierarchy is constructed explicitly during the setup phase (Algorithm 5.2). Most of the current methods can be classified in terms of coarsening schemes as *classical* and *aggregation* AMG. The classical AMG [42, 212, 247] can be considered as an algebraic counterpart of the traditional geometric multigrid (used in Chapter 2). A certain subset of the fine-level variables is identified with the variables on the coarse level and a linear interpolation operator is deduced from the matrix entries to approximate the values on the fine points from the values on the neighbouring coarse points.

Algorithm 5.1 Recursive multigrid V-cycle $\mathrm{MG}_\ell(f_\ell, v_\ell)$

Input: Assume that the following is given

- $A_\ell \in \mathbb{R}^{n_\ell \times n_\ell}$ ▷ Operators discretized on Ω_l $(\ell = 1, \cdots, L)$
- $f_\ell \in \mathbb{R}^{n_\ell}$ ▷ Right-hand side given on Ω_l
- $v_\ell \in \mathbb{R}^{n_\ell}$ ▷ Initial guess given on Ω_l

if $\ell = L$ **then**
  solve $A_\ell v_\ell = f_\ell$
else
  $v_\ell \leftarrow S_{\ell,\mathrm{pre}}(f_\ell, v_\ell)$ ▷ Pre-smoothing
  $r_{\ell+1} \leftarrow R_\ell(f_\ell - A_\ell v_\ell)$ ▷ Residual restriction
  $c_{\ell+1} \leftarrow 0$ and do $\mathrm{MG}_{\ell+1}(r_{\ell+1}, c_{\ell+1})$ ▷ Coarse-grid correction
  $v_\ell \leftarrow v_\ell + P_\ell c_{\ell+1}$ ▷ Solution update
  $v_\ell \leftarrow S_{\ell,\mathrm{post}}(f_\ell, v_\ell)$ ▷ Post-smoothing
end if

In the aggregation methods [36, 38, 51, 191], the coarse-level variables are associated with contiguous disjoint groups of fine-level variables called aggregates. The interpolation is defined simply by an injection of the values from the coarse-level to the fine-level variables in each aggregate.

Algorithm 5.2 AMG setup

Input: Assume that the following is given

- $A_\ell \in \mathbb{R}^{n_\ell \times n_\ell}$ ▷ Operators discretized on Ω_l $(\ell = 1, \cdots, L)$
- $f_\ell \in \mathbb{R}^{n_\ell}$ ▷ Right-hand side given on Ω_l
- $v_\ell \in \mathbb{R}^{n_\ell}$ ▷ Initial guess given on Ω_l

1: set $\Omega_1 \leftarrow \Omega$, $A_1 \leftarrow A$, $\ell = 1$
2: **while** Ω_ℓ is not small enough and $\ell < L_{\max}$ **do**
3: Construct the coarse-level variables $\Omega_{\ell+1}$
4: Compute the interpolation operator P_ℓ
5: Set $R_\ell = P_\ell^T$
6: Compute the coarse-level operator $A_{\ell+1} = R_\ell A_\ell P_\ell$
7: $\ell \leftarrow \ell + 1$
8: **end while**

The AMG methods based on the classical approach usually exhibit very good performance in terms of the convergence rates of preconditioned iterative methods (that is, the iteration counts to achieve a certain stopping criterion). The price to pay is however sometimes characterized by quite high computational costs and memory requirements associated with the setup and solution phases of the preconditioner. The aggregation AMG methods, on the other hand, are often relatively cheaper with low operator complexities. Nevertheless, they are rarely used in the multigrid setting because of the fact that the convergence rate of an aggregation multigrid method often depends on the problem size [38, 247, 264], that is, on the number of levels. The scalability of the aggregation methods can be improved, for example, by "enriching" the interpolation operators [44, 268] or by using the over-correction [38] or a certain polynomial acceleration [16, 191, 193] of the coarse-level correction step. We note that the convergence of aggregation methods can be made independent of the problem size for the two-level method or, more generally, when the number of levels is limited; see, for example, [36, 187]. Nevertheless this requires solving possibly large coarse systems to a high level of accuracy, for example, by using a direct solver, which can be impractical in practice.

Here, we consider combining the aggregation and classical AMG approaches within one hierarchy or, equivalently, replacing the "exact" coarse-level solver in the aggregation part of the hierarchy by a classical AMG solver. Thus we aim at obtaining *reduced* costs for the setup and solve phases of the resulting preconditioner, while retaining the optimal convergence of the "exact" aggregation method without the need to solve the coarse level problems to a high level of accuracy.

Collaboration

This research project was initially developed with Serge Gratton, Pascal Hénon (TOTAL) and Pavel Jiránek with financial support from TOTAL.

Problem setting

The classical AMG method with a short-distance interpolation as a preconditioner (or stationary solver) often leads to very good V-cycle convergence rates for a wide range of problem sizes. However, the operator complexities characterizing approximately the memory requirements and computational costs of the V-cycle (when ignoring the costs associated with the inter-level operators), defined by

$$C_{op} = \frac{1}{\mathrm{nz}(A_1)} \sum_{\ell=1}^{L} \mathrm{nz}(A_\ell), \tag{5.1}$$

where $\mathrm{nz}(A_\ell)$ denotes the number of nonzero entries in the operator A_ℓ on the level ℓ, can be relatively high. On the other hand, the plain aggregation AMG is relatively fast and cheap but if the number of levels is not kept fixed independently of the problem size, the convergence rate of the V-cycle (even of the W-cycle) may deteriorate quickly for increasingly large problems.

It is possible to reduce the complexity of the AMG method by first performing the aggressive coarsening with a long-range interpolation [247] on a prescribed number of fine levels and then applying a less aggressive coarsening with a more accurate short-distance interpolation to create the remainder of the multigrid hierarchy. Such an approach is also implemented in the BoomerAMG package of the Hypre library [145]. Although the convergence of preconditioned iterative methods often deteriorates, the substantial decrease of the operator complexity may compensate for this and, consequently, the combination of these two approaches in one AMG hierarchy can provide a significant performance improvement.

Motivated by this idea, we consider replacing the classical aggressive coarsening approach by aggregation. The resulting AMG method combines a fixed number of (plain) aggregation levels with the rest of the hierarchy generated by a classical AMG scheme with a short-range interpolation. Alternatively, this can be considered as using an aggregation AMG method (with a limited number of levels) with the coarse-level solver replaced by the classical AMG. We refer the reader to [132] for further details on the algorithmic part.

Application and future tasks

We briefly investigate the weak scalability properties of the new multigrid variants and then describe future lines of research.

Application This research is motivated by the particular application of AMG in the context of solving linear algebraic systems arising from multiphase flow models in reservoir simulations. Discretization and linearization of such models lead to a large sequence of systems, which couple together the pressure and saturation or concentration unknowns in the computational cells of the underlying domain discretization and which must be solved efficiently. A possible approach for preconditioning such systems is based on decoupling the dependency of the pressure on the remaining unknowns. An efficient solution method is thus required for the linear systems related to the pressure variables. This is investigated next.

We denote by AGGnRr the variant based on n levels of plain agglomeration followed by the classical AMG coarsening procedure, r being a parameter controlling the initial coarsening ratio. We compare these variants with aggressive AMG on the first level (AGGRES1) or on the first two levels (AGGRES2) followed by classical AMG coarsening in the hierarchy.

We evaluate the weak scalability properties of the multigrid variants on an anisotropic Poisson problem defined on the three-dimensional unit cube with homogeneous Dirichlet boundary conditions. The fine level operator is discretized by the 7-point finite difference stencil on a uniform regular grid with $64^3 = 262144$ grid points per MPI process. The AMG methods are used as fixed preconditioners for the conjugate gradient method and the iterations are stopped when the relative residual norm decreases below the tolerance 10^{-6}. Table 5.1 collects numerical results related to these weak scalability experiments. Whatever the number of MPI processes, we observe that the minimal

Method	$\#MPI$	Setup	Solve	Total	C_{op}	L	It
AGG1R1	1	3.773e-01	5.141e-01	*8.914e-01*	2.193	7	12
	27	1.345e+00	8.994e-01	2.244e+00	2.196	9	14
	125	1.842e+00	1.120e+00	2.962e+00	2.244	9	16
	343	2.024e+00	1.268e+00	*3.292e+00*	2.258	10	16
	729	2.625e+00	1.421e+00	4.046e+00	2.264	10	17
AGGRES1	1	4.760e-01	5.452e-01	1.021e+00	2.444	6	10
	27	1.728e+00	1.166e+00	2.894e+00	2.322	8	15
	125	2.113e+00	1.610e+00	3.723e+00	2.436	10	15
	343	2.452e+00	1.523e+00	3.975e+00	2.490	10	16
	729	3.403e+00	1.580e+00	4.983e+00	2.522	10	16
AGG2R1	1	3.133e-01	5.903e-01	9.036e-01	1.784	7	17
	27	8.780e-01	1.179e+00	*2.057e+00*	1.836	8	23
	125	1.154e+00	1.389e+00	*2.543e+00*	1.863	9	25
	343	2.214e+00	1.551e+00	3.765e+00	1.873	9	26
	729	1.922e+00	1.763e+00	*3.685e+00*	1.879	10	27
AGGRES2	1	3.525e-01	6.646e-01	1.017e+00	1.833	5	14
	27	1.113e+00	1.333e+00	2.446e+00	1.748	7	20
	125	1.448e+00	1.555e+00	3.003e+00	1.796	8	21
	343	2.028e+00	1.899e+00	3.927e+00	1.854	8	22
	729	2.973e+00	2.105e+00	5.078e+00	1.887	9	22

Table 5.1.: Weak scalability experiments for the anisotropic Poisson problem (executed on up to 729 MPI processes, each process using 8 tasks). Setup and Solve correspond to the computational time spent in the setup and solution phases, respectively. C_{op} denotes the operator complexity (relation (5.1)), L the total number of levels in the hierarchy, It the number of conjugate gradient iterations required to decrease the residual norm by 6 orders of magnitude. Table III of [132].

computational times (including both setup and solution phases) are related to either AGG1R1 or AGG2R1, i.e., the new variants that have been proposed (see italic entries in Table 5.1). Most of the performance gain with respect to existing efficient schemes in the literature (AGGRES1 and AGGRES2) is indeed due to the simplicity of the aggregation operators: both their construction and their application require no communication at all. Consequently, applying a cycle related to the AGG1R1 or AGG2R2 variants will be cheap both in terms of computations and communications. This fact has been also confirmed on realistic three-dimensional problems related to reservoir modelling (SPE10 and VISFIN3D problems [132, Section 4]), where the efficiency of these variants has been demonstrated as well. We also refer the reader to [132, Section 4] for additional comments related to strong scalability experiments. Finally, the full hierarchy information related to the AGG2R1 variant is given in Table 5.2 in the case of 729 MPI processes.

Method	Level	$\#Rows$	$\#Nz$	s_{min}	s_{max}	s_{avg}
AGG2R1	1	191102976	1335730176	4	7	7.0
	2	65691648	458721792	4	7	7.0
	3	23887872	166385664	4	7	7.0
	4	11612160	170627328	6	15	14.7
	5	5579438	275592862	16	101	49.4
	6	961931	93306855	26	153	97.0
	7	92607	9160225	24	159	98.9
	8	12964	916080	17	142	70.7
	9	949	10781	1	28	11.4

Table 5.2.: Hierarchy information for the AGG2R1 variant applied to the anisotropic Poisson problem (executed on 729 MPI processes, each process using 8 tasks). $\#Rows$ denotes the number of rows in the matrix, $\#Nz$ the total number of nonzero entries. s_{min} and s_{max} denote the minimum and maximum numbers of nonzero entries per row, respectively. s_{avg} corresponds to the total number of nonzero entries divided by the total number of rows. Information extracted from Table V of [132].

Future tasks

- **Reuse of information in AMG for sequence of linear systems**.

  As previously discussed in Chapter 4, we are often faced with the solution of linear systems coming from the discretization of elliptic partial differential equations with variable coefficients. In such a setting, we aim at deriving a *scalable* solution method not only for a single linear system but for the *whole* sequence. The key idea is thus to reduce the computational times further by reusing information obtained during the previous setup and/or solution phases of the algebraic multigrid method. Two possible situations can be described in more detail.

  Sequences with a fixed left-hand side matrix and with multiple changing left-hand sides. We consider here the case of a fixed matrix and right-hand sides that may be given either in sequence or simultaneously. Consequently, the setup phase has to be performed only once and the complete multilevel hierarchy (including transfer operators and coarse level operators) can be reused over the whole sequence. Moreover, when the right-hand sides are given simultaneously, we note that a block version of the algebraic multigrid method has to be favoured to take advantage of efficient BLAS-3 kernels (matrix-matrix operations) as much as possible. The solution phase can be even further improved by using recycling subspace techniques or by updating the preconditioners to further accelerate the convergence of the multigrid variant; both strategies are discussed later in Section 5.4.1.

  Sequences with changing matrices. This situation is quite relevant, since it is usually the most frequent one. We note that considering partial differential equations with different sets of coefficients (as in uncertainty quantification) or solving non-

linear problems with Picard or Newton methods correspond to such a setting. This latter situation has been investigated in [151] and in [131] with the new variant of AMG in the context of reservoir simulations, assuming that the fine level matrices in the sequence share the same sparsity pattern (i.e. connectivity graph). Given this assumption, the transfer operators related to the agglomeration procedure are then identical over the whole sequence. Consequently, the sparsity patterns of the corresponding coarse level operators are identical. Hence, it is thus only necessary to compute the nonzero entries of such operators (matrix-matrix-matrix operations). Reduced computational times for the set-up phase have been obtained in such a setting. Improvements related to the solution phase can be obtained as well by using the same techniques as described above. This needs to be further investigated.

- **Software**.

  A public domain implementation of the proposed algebraic multigrid method is targeted on a long-term basis. Its core will be the PMG code developed by Pavel Jiránek, when he was at CERFACS. Additional kernels will be proposed to take into account the case of sequences of linear systems or more generally, a family of linear systems. Based on this software, weak and strong scalability performance studies for sequences of linear systems will be performed on both academic and realistic applications. First, this package will address the solution of linear systems with symmetric positive definite matrices. In addition, we aim to provide an implementation that authorizes the solution of linear systems with saddle point structure. In such a setting, we assume that the $(1,1)$ block of the global saddle point matrix is symmetric positive definite. Block triangular preconditioners [154] or Schur complement preconditioners [90] will be then employed, where a cycle of the multigrid preconditioner will be used as an approximate inverse of the $(1,1)$ block. Applications to fluid mechanics, geophysics and structural mechanics are anticipated.

5.3.2. Combination of multilevel domain decomposition and algebraic multigrid methods

Motivations

Two-level domain decomposition methods provide *scalable* preconditioners with respect to the number of subdomains for self-adjoint elliptic partial differential equations [246]. Non-overlapping iterative substructuring methods such as FETI and BNN are typical examples, as detailed in Chapter 3 in the context of *hp* finite element approximations. Two-level overlapping preconditioners have been also proposed in the literature; see, e.g., [246] and the recent monograph [80]. To address the question of scalability for the solution of large-scale linear systems on forthcoming computers, it is thus necessary to also rely on efficient and scalable solvers for the local problems defined on each subdomain and for the global coarse problem, respectively. This key issue is addressed next in this section.

Collaboration

This project has emerged while preparing the SOLEX research proposal coordinated by Santiago Badia (Universitat Politècnica de Catalunya and CIMNE, Spain). We are currently looking for financial support to start part of this activity.

Problem setting

We propose to combine non-overlapping domain decomposition and algebraic multigrid methods to address this question. This combination is however not new; see [246, Section 4.3] for early references, where a cycle of multigrid is used as a local solver. We first focus on the choice of the non-overlapping domain decomposition preconditioner. Interesting candidates are quite recent methods known as BDDC (Balancing Domain Decomposition by Constraints) [79] and FETI-DP (Dual Primal Finite Element Tearing and Interconnecting) [102]. Indeed, FETI-DP and BDDC methods [173] share the following salient properties

- Coarse and local components can be computed in a parallel additive way; this has been exploited in [18, 19, 20].
- Both methods allow for an extremely aggressive coarsening. Indeed, the coarse matrix has a similar sparsity pattern to the original fine level matrix. This feature is really essential in terms of operator complexity. This appears as a major difference compared with the class of algebraic multigrid methods, where the coarse matrices constructed with the Galerkin method exhibit larger stencils versus the number of levels in the hierarchy (see Table 5.2 for an illustration on an academic problem).
- Both methods allow the use of inexact methods as local solvers or global coarse solver, without impacting on the convergence rate of the preconditioned method [20, 157, 173, 206].
- Multilevel extensions of FETI-DP and BDDC methods have been proposed in the literature [174].

Future tasks and applications

We propose to use the algebraic multigrid method presented in Section 5.3.1 as a local solver on the subdomains within the BDDC or FETI-DP algorithms to address the solution of self-adjoint elliptic partial differential equations. In this context, we note that the local subdomain matrices are symmetric positive definite; the algebraic multigrid method proposed in Section 5.3.1 can then be applied straightforwardly. We anticipate that this combination will offer these promising features

- both the weak and strong scalabilities of such domain decomposition methods will be improved, since an efficient iterative method will replace sparse direct methods that usually exhibit larger operator complexities,

- the global memory requirements of the domain decomposition methods will be drastically reduced, since local iterative solvers are now employed.

Both features are especially important when tackling large-scale applications on massively parallel computers. Finally, we would like to mention that overlapping two-level domain decomposition methods may benefit from this algebraic multigrid solver as well. Interesting recent candidates are based on Restricted Additive Schwarz methods [56] with local Dirichlet-to-Neumann operators [67, 68, 188] or the GenEO method (Generalized Eigenvalue problems on the Overlap) to compute an adaptive coarse space [150, 230].

Software Open-source domain decomposition software that could be considered in this setting are, e.g., BDDCML[1], FEMPAR[2] or HPDDM[3].

Applications Targeted applications are related to the solution of self-adjoint elliptic partial differential equations with variable coefficients that arise, e.g., in geoengineering sciences (subsurface flow simulations in porous media, oil and gas reservoirs, pollutant transport, nuclear waste deposits) or in structural mechanics.

5.4. Scalable algorithms beyond linear solvers

We next present two research prospectives related to specific settings of interest, when simulating physical phenomena based on partial differential equations. When available, numerical results are presented.

5.4.1. Sequences of systems

Motivations

Sequences of systems often occur when considering the solution of optimization problems, nonlinear deterministic or stochastic partial differential equations or eigenvalue problems, to name a few. Efficient numerical methods must then be designed for such a purpose. When the right-hand sides are available simultaneously, preconditioned block Krylov subspace solvers have proved efficient and useful (see Chapter 4), when the dimension of the problem is large and/or when the preconditioner application is known to be expensive. When the right-hand sides are given in sequence, recycling Krylov subspace methods have been proposed in the literature; see, e.g., [119, 120, 197]. The main idea is to reuse subspace information to improve the convergence rate of the Krylov subspace method when solving the subsequent linear systems. FGCRO-DR [61], also presented in Section 4.3.6, belongs to such a family. Here, we propose to follow a different and complementary path by improving the preconditioner using techniques

[1] http://users.math.cas.cz/~sistek/software/bddcml.html
[2] https://web.cimne.upc.edu/groups/comfus/fempar.html
[3] https://github.com/hpddm/hpddm

coming from numerical optimization. The main goal is to improve an already existing preconditioner (referred to as a first level preconditioner) by exploiting any available information (that can be cheaply obtained) to approximate the matrix inverse with respect to a certain subspace. This strategy, known as Limited Memory Preconditioner (LMP) in the numerical optimization literature [182, 190], has been analysed in the symmetric positive definite case [128]. Our first objective is to extend this idea to the case of a sequence of linear systems with symmetric indefinite matrices.

Collaboration

This research project was initially developed with Serge Gratton, Sylvain Mercier and Nicolas Tardieu (EDF) with financial support from ANRT (Association Nationale de la Recherche et de la Technologie). Part of the current prospectives is developed within the PAMSIM project funded by Bpifrance (Banque Publique d'Investissement).

Problem setting

We address the solution of a sequence of linear systems of the form

$$A\ x_i = b_i, \quad i = 1, \ldots, I,$$

with $A \in \mathbb{R}^{N\times N}$ being a large symmetric indefinite matrix, $x_i \in \mathbb{R}^N$ and $b_i \in \mathbb{R}^N$. We assume that a first level preconditioner is already available so that the matrix A corresponds to a preconditioned operator in our setting. Our main interest will be to analyse the class of limited memory preconditioners defined next.

Definition 5.1. Let A be a symmetric indefinite matrix of order N. Assume that $S \in \mathbb{R}^{N\times k}$, with $k \leq N$, is such that S^TAS is nonsingular. The symmetric matrix H defined as

$$H = (I_N - S(S^TAS)^{-1}S^TA)(I_N - AS(S^TAS)^{-1}S^T) + S(S^TAS)^{-1}S^T \tag{5.2}$$

is called the limited memory preconditioner in the indefinite case.

Properties of this preconditioner have been analysed in [130]. A formula to characterize the spectrum of the preconditioned operator has been derived (see Theorem 3.2 of [130]). We have shown that the eigenvalues of the preconditioned operator are real-valued (with at least k eigenvalues equal to 1). Furthermore, we have shown that the eigenvalues of the preconditioned matrix enjoy interlacing properties with respect to the eigenvalues of the original matrix provided that the k linearly independent vectors have been previously projected onto the invariant subspaces associated with the eigenvalues of the original matrix in the open right and left half-plane, respectively. This main result is stated hereafter. These projection operators involving the matrix sign function of A are defined next [147].

Definition 5.2. Let $A \in \mathbb{R}^{N\times N}$ be a symmetric indefinite matrix of order N and let $X \in \mathbb{R}^{N\times N}$ denote the matrix sign function[4] of A defined as $X = (A^2)^{-\frac{1}{2}}A$. Let $\mathcal{I}_+(A)$ and $\mathcal{I}_-(A)$ denote the invariant subspaces associated with the eigenvalues in the right and left half-plane, respectively. We define $P_+(A) = (I_N + X)/2$ as the projection operator onto $\mathcal{I}_+(A)$ and $P_-(A) = (I_N - X)/2$ as the projection operator onto $\mathcal{I}_-(A)$, respectively.

We denote by $Q_+ \in \mathbb{R}^{N\times N_+}$ ($Q_- \in \mathbb{R}^{N\times N_-}$) an orthonormal basis of $\mathcal{I}_+(A)$ ($\mathcal{I}_-(A)$, respectively) and by $Q \in \mathbb{R}^{N\times N}$ the orthonormal matrix defined as $Q = [Q_+, Q_-]$ with $N = N_+ + N_-$. Given $\tilde{S} \in \mathbb{R}^{N\times k}$, $S = [S_+, S_-]$ ($S_+ \in \mathbb{R}^{N\times k_+}$, $S_- \in \mathbb{R}^{N\times k_-}$ with $k = k_+ + k_-, k \leq N$) consists of k projected vectors obtained as

$$S_+ = Q_+Q_+^T[\tilde{s}_{i_1}, ..., \tilde{s}_{i_{k_+}}], \tag{5.3}$$

$$S_- = Q_-Q_-^T[\tilde{s}_{j_1}, ..., \tilde{s}_{j_{k_-}}], \tag{5.4}$$

where $[\tilde{s}_{i_1}, ..., \tilde{s}_{i_{k_+}}]$ ($[\tilde{s}_{j_1}, ..., \tilde{s}_{j_{k_-}}]$) corresponds to k_+ (k_-, respectively) distinct columns of $\tilde{S}$. Equivalently, we can write

$$S_+ = Q_+\,\tilde{S}_+, \quad \tilde{S}_+ \in \mathbb{R}^{N_+\times k_+}, \quad \tilde{S}_+ = Q_+^T[\tilde{s}_{i_1}, ..., \tilde{s}_{i_{k_+}}] \tag{5.5}$$

$$S_- = Q_-\,\tilde{S}_-, \quad \tilde{S}_- \in \mathbb{R}^{N_-\times k_-}, \quad \tilde{S}_- = Q_-^T[\tilde{s}_{j_1}, ..., \tilde{s}_{j_{k_-}}]. \tag{5.6}$$

The main goal is to show that a property of nonexpansion of the spectrum of HA can be obtained by solving two tractable subproblems related to either $\mathcal{I}_+(A)$ or $\mathcal{I}_-(A)$. $\mathcal{I}_+(A)$ and $\mathcal{I}_-(A)$ are H-invariant, see Lemmas 3.6 and 3.7 in [130]. We define $A_+ = Q_+^TAQ_+ \in \mathbb{R}^{N_+\times N_+}$ ($A_- = Q_-^TAQ_- \in \mathbb{R}^{N_-\times N_-}$) as the orthogonally projected restriction of A with respect to the basis Q_+ (Q_-, respectively) and $H_+ = Q_+^THQ_+ \in \mathbb{R}^{N_+\times N_+}$ ($H_- = Q_-^THQ_- \in \mathbb{R}^{N_-\times N_-}$) the orthogonally projected restriction of H with respect to the basis Q_+ (Q_-, respectively). The main theorem is stated next.

Theorem 5.3. *Let A be a symmetric indefinite matrix of order N, H be given by (5.2) in Definition 5.1 based on $S = [S_+, S_-]$ consisting of k_+ (k_-) vectors projected onto the positive (negative, respectively) invariant subspace of A, $\mathcal{I}_+(A)$ ($\mathcal{I}_-(A)$, respectively). Then, the following properties hold*
(a) Let the positive real numbers $\sigma_1^+, \cdots, \sigma_{N_+}^+$ denote the eigenvalues of A_+ sorted in nondecreasing order. Then the set of eigenvalues $\mu_1^+, \cdots, \mu_{N_+}^+$ of H_+A_+ can be split in two subsets

$$\begin{aligned} \sigma_j^+ \leq \mu_j^+ \leq \sigma_{j+k_+}^+ \text{ for } j \in \{1, \cdots, N_+ - k_+\}, \\ \mu_j^+ = 1 \text{ for } j \in \{N_+ - k_+ + 1, \cdots, N_+\}. \end{aligned} \tag{5.7}$$

(b) Let the negative real numbers $\sigma_1^-, \cdots, \sigma_{n_-}^-$ denote the eigenvalues of A_- sorted in nondecreasing order. Then the set of eigenvalues $\mu_1^-, \cdots, \mu_{N_-}^-$ of H_-A_- can be split in two subsets

$$\begin{aligned} \sigma_j^- \leq \mu_j^- \leq \sigma_{j+k_-}^- \text{ for } j \in \{1, \cdots, N_- - k_-\}, \\ \mu_j^- = 1 \text{ for } j \in \{N_- - k_- + 1, \cdots, N_-\}. \end{aligned} \tag{5.8}$$

[4] A (being symmetric indefinite) has no eigenvalues on the imaginary axis, so that the matrix sign function of A is defined.

(c) In addition, the condition number of HA, $\kappa(HA)$, can be bounded as follows

$$\kappa(HA) \leq \frac{\max\{1, \sigma_{N_+}^+, |\sigma_1^-|\}}{\min\{1, \sigma_1^+, |\sigma_{N_-}^-|\}}. \tag{5.9}$$

Proof. See Theorem 3.10 of [130]. □

As stated in Theorem 5.3, the use of projected vectors in the Limited Memory Preconditioner ensures a nonexpansion property of the spectrum of the preconditioned operator, which is an attractive feature. Nevertheless, using the exact sign function of A or matrix functions that approximate $\text{sign}(A)\tilde{S}$ can be computationally too expensive for large-scale problems. Consequently, approximate spectral information based on Ritz vectors (information that is cheaply available) is usually chosen to select the k columns of $\tilde{S}$. This leads to the Ritz Limited Memory Preconditioner (Ritz-LMP). We refer the reader to Section 3.5 of [130] for a theoretical analysis of Ritz Limited Memory Preconditioner. There, a discussion on practical aspects related to computational cost and memory requirements in such a setting is also provided.

Applications and future tasks

An application to solid and structural mechanics is briefly considered next, where efficient preconditioners for linear systems with saddle point structure must be proposed. Here we investigate the efficiency of a Ritz Limited Memory Preconditioner and refer the reader to [130, 179] for further details on the context.

Application to structural mechanics: containment building of a nuclear reactor In such an application, we consider a sequence of linear systems of saddle point type

$$\mathcal{K}_i \; y_i = c_i \Longleftrightarrow \begin{pmatrix} G_i & B^T \\ B & 0_{m,m} \end{pmatrix} \begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} f_i \\ g_i \end{pmatrix}, \quad i = 1, \cdots, I, \tag{5.10}$$

where $G_i \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$, $f_i \in \mathbb{R}^n$, $g_i \in \mathbb{R}^m$ and $m < n$ (hence $N = m + n$). u_i denote the physical unknowns and v_i the Lagrange multipliers. The stiffness matrices G_i ($i = 1, \cdots, I$) are symmetric positive semidefinite since they are related to the discretization of an unconstrained mechanical problem (i.e. with no essential boundary conditions). The deficiency of G_i can be large. Indeed it is known that an upper bound of the dimension of $\mathcal{N}(G_i)$ corresponds to the number of rigid body motions of subbodies of materials contained within the finite element mesh. Here these motions correspond to three translations and three rotations for each subbody [189]. We further assume that B is of full row rank ($rank(B) = m$) and that $\mathcal{N}(G_i) \cap \mathcal{N}(B) = \{0\}$, $\forall i \in \{1, \cdots, I\}$. These assumptions ensure the existence and uniqueness of the solution of each linear system in the sequence [28]. We also note that B is a very sparse matrix in our setting. Indeed, B is usually related to the dualization of the boundary conditions. These relations are local in the sense that they involve adjacent nodes of the mesh. Unless stated, B admits only one nonzero coefficient per row due to Dirichlet boundary conditions. In this case, $B^T B$ is a diagonal matrix.

An approximate factorization of a block diagonal symmetric positive definite preconditioner [205] is considered as a first level preconditioner [130] and we later analyse the effect of the Ritz-LMP preconditioner for the solution of the sequence of linear systems. The LMP preconditioner is based on Ritz information obtained at the end of the solution phase of the first linear system. The same preconditioner is used through the sequence. We consider an industrial geometry proposed by EDF known as a containment building of a nuclear reactor (see Figure 5.1), for which numerical simulations are performed to produce a safety analysis. We refer the reader to [130] for a complete description of the physical context and comment here only on aspects related to the proposed preconditioner.

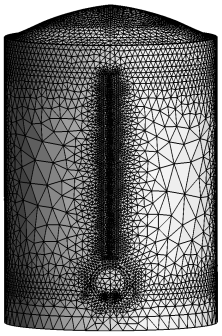


Figure 5.1.: Containment building: three-dimensional mesh. Figure 4.4 of [130].

Figure 5.2 shows the evolution of the Euclidean norm of the relative residual for the last three linear systems in the sequence ($I = 2, 3, 4$). In this experiment, we consider limited memory preconditioners with a varying number of Ritz vectors ($k = 5, 20, 30$, respectively). Whatever the linear system considered in the sequence, the smallest number of iterations is obtained when selecting a large value of Ritz vectors ($k = 30$). In addition, we show in Table 5.3 the cumulative iteration count over the last three linear systems, the total number of floating-point operations (one floating-point operation corresponding to one real number operation of multiply/divide/add/subtract type) and the memory requirements, respectively. We note that selecting S based on $k = 30$ Ritz vectors leads to a decrease of 47% in terms of cumulative iteration count and to a decrease of 43% in terms of computational operations. This satisfactory result comes at a price of a very moderate increase in memory requirements (3%), since the limited memory preconditioner only needs the storage of $(k + 2)$ vectors of size N.

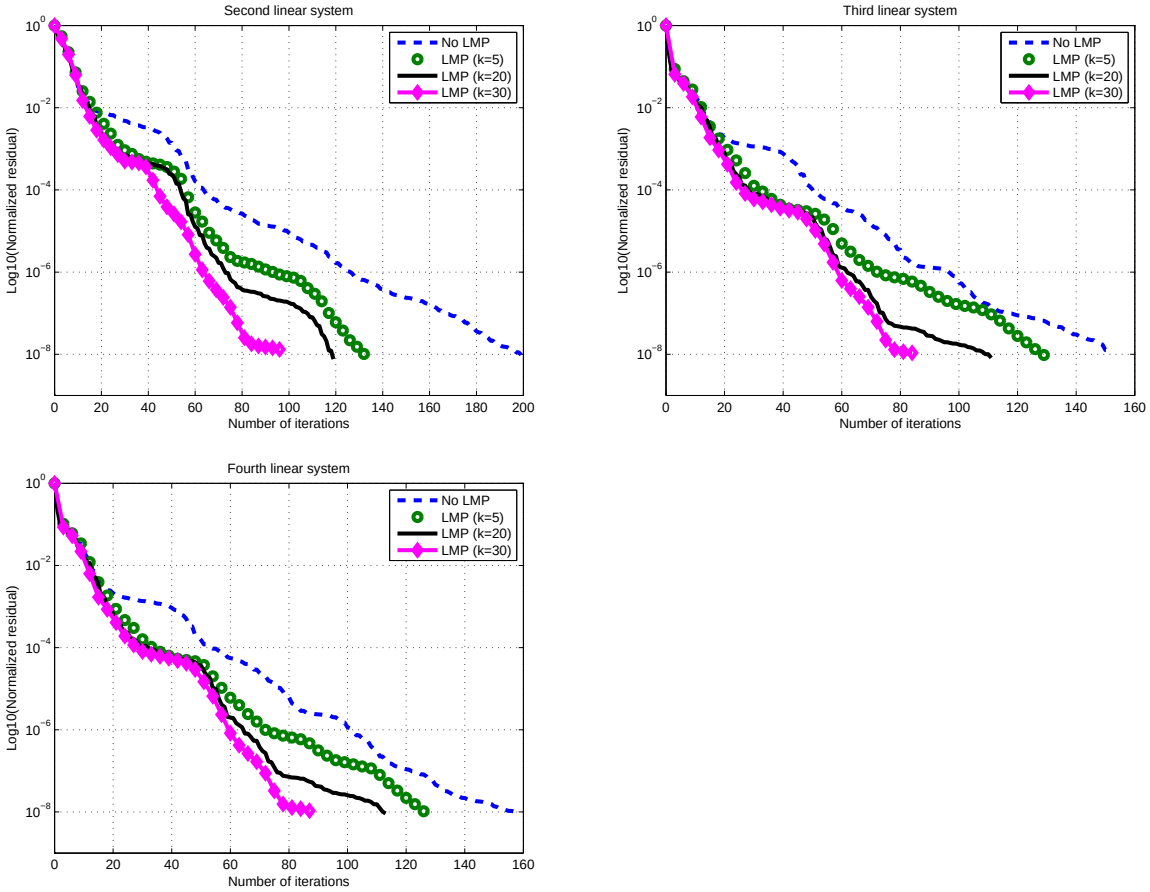


Figure 5.2.: Containment building: convergence history of preconditioned GMRES(30) for the last three linear systems in the sequence. Case of limited memory preconditioners with $k = 5$, 20 or 30 Ritz vectors associated to the smallest in modulus Ritz values. Figure 4.5 of [130].

	No LMP$_{\pm}$	LMP$_{\pm}$, $k = 5$	LMP$_{\pm}$, $k = 20$	LMP$_{\pm}$, $k = 30$
Total iteration count	509	389	343	272
Iteration count decrease (%)	×	24	33	**47**
CPU time (sec)	315	254	224	186
CPU time decrease (%)	×	19	29	**41**
Memory (Mo)	6686	6722	6823	6891
Memory increase (%)	×	**0.5**	2	3

Table 5.3.: Containment building: cumulative iteration count for the last three linear systems in the sequence, CPU time and memory requirements for different limited memory preconditioners. Case of $k = 5$, 20 or 30 Ritz vectors. Table 4.1 of [130].

Future tasks

- **Extension to the nonsymmetric case**

  We aim at proposing and analysing preconditioner update formulae to be used in a more general setting, when either the matrix A or the first-level preconditioner is nonsymmetric. Possible candidates are

  $$H_{\pm} = (I_N - S(S^T A^T S)^{-1} S^T A^T)(I_N - AS(S^T AS)^{-1} S^T) + S(S^T AS)^{-1} S^T, \quad (5.11)$$

  where $S \in \mathbb{R}^{N \times k}$ is such that $S^T AS$ is nonsingular, or

  $$H_{ns} = (I_N - AS(S^T A^T AS)^{-1} S^T A^T) + S(S^T A^T AS)^{-1} S^T A^T, \quad (5.12)$$

  where $S \in \mathbb{R}^{N \times k}$ is of full rank k. H_{ns} has been notably used in [179] and successfully applied to problems in structural mechanics. H_{ns} has been derived from variants proposed by Broyden [47, 48] and Eirola and Nevanlinna [88], respectively (see Chapter 3 of [179] for more details). The analysis of the mathematical properties of $H_{\pm}$ and H_{ns} remains to be performed in this setting. We note that simpler expressions of $H_{\pm}$ and H_{ns} can be obtained, when S is based on Ritz and harmonic Ritz information, respectively. This leads to attractive update formulae, since they are particularly cheap to apply to a given vector.

- **Selection of deflation vectors** S

  As described in Section 5.4.1, deflation vectors have been selected according to the k smallest in magnitude Ritz values with k equal to 5, 20 or 30. Heuristics are usually employed to select the deflation vectors, as frequently described in the literature. Automatic selection of deflation vectors subject to certain constraints (e.g. the global cost of the preconditioner application and their impact on the convergence rate of the Krylov subspace method) would be nice to perform. Promising attempts have been proposed in both the symmetric positive definite and symmetric indefinite cases in [119, 120], respectively. Extensions to the nonnormal case are required to tackle current situations of interest. Nevertheless, we know that general statements about the optimal choice of deflation vectors seem unrealistic, unless the convergence behaviour of the Krylov subspace method is better understood. Moreover, in our context, we need to consider Krylov subspace methods that allow variable preconditioners, for which considerably less theory has been proposed. Additional theory is thus required to fully understand how to choose the deflation vectors properly and to understand the links with deflated Krylov subspace solvers.

The above future tasks have exclusively concerned the analysis of the limited memory preconditioner to address the solution of linear systems given in sequence. We close this section with open questions related to the mathematical analysis of methods for sequences of linear and nonlinear systems. We would like to address these following points more specifically

- **Mathematical analysis of recycling methods**

  How does the known behaviour of a Krylov subspace method applied to $Ax = b$ relate to the behaviour of the same Krylov subspace method applied to $(A+E)y = b + e$, where both E and e are matrix- and vector-perturbations, respectively ? This question has been already addressed in the literature [119, Section 3.4] but to the best of our knowledge, remains to be investigated for flexible Krylov subspace methods with general A.

- **Choice of the deflation subspace**

  The second question is related to deflation applied to a sequence of linear systems. For the sake of clarity and brevity, we consider a sequence made of only two linear systems that are supposed to share some common properties (i.e. may be from the discretization of the same nonlinear problem). Let us assume that the first linear system $Ax = b$ has been solved with a deflated Krylov subspace method with deflation subspace $\mathcal{U}$. As quoted in [121, 137], this is equivalent to the solution of the deflated linear system $\mathsf{P}Ax = \mathsf{P}b$ where P is a projection operator based on $\mathcal{U}$. The main question to be answered is how to find an effective deflation subspace $\mathcal{V} \subset \mathcal{K}(\mathsf{P}A, \mathsf{P}b)$ to be applied to the next linear system $By = c$ (with an oblique or orthogonal projection based on the new deflation subspace $\mathcal{V}$) ?

- **Differentiating Krylov subspace methods**

  For ease of exposition, we consider only two linear systems in a given sequence, say $Ax = b$ and $Ay = \widehat{c}$ as the first and second linear systems, respectively. Let denote x_k the kth iterate of a Krylov subspace method when solving $Ax = b$. For fixed A, initial guess x_0 and k, the iterate $x_k = x_k(b)$ is a nonlinear function of b. It can be verified that x_k is also a differentiable function of b and expressions for the Jacobian of x_k with respect to the right-hand side vector b, i.e.,

  $$J_k = \frac{\partial x_k}{\partial b} \tag{5.13}$$

  can be obtained (see [129] in the case of the method of conjugate gradients). If ℓ is the number of iterations required to solve the first linear system to a given accuracy, $x_l(b)$ the corresponding approximation solution of the first linear system, the information about the Jacobian can be used to easily derive an initial guess for the next linear system in the sequence using the first-order expansion

  $$x_0(c) = x_\ell(b) + J_\ell(c - b). \tag{5.14}$$

  We refer the reader to [129, Section 5.5] where encouraging results have been obtained in the symmetric positive definite case. Extension to general linear systems with GCR [89] as a Krylov subspace method is currently investigated.

- **Nonlinear acceleration techniques**

Physical phenomena are often governed by nonlinear partial differential equations. Newton-Krylov methods [75] are usually employed as nonlinear solvers in application codes in industry, since they may be easily built from existing linear solvers and preconditioners (including multilevel preconditioners such as geometric multigrid or domain decomposition methods). However, a possible lack of robustness occurs when the initial guess is far from the solution. Thus there is probably a large design space related to nonlinear solvers to improve or even replace the Newton method. Nonlinear multigrid methods such as FAS [39] or nonlinear additive Schwarz method [55] represent two such alternatives. It is also of primary interest to design efficient nonlinear solvers and/or nonlinear acceleration techniques that exhibit sufficient arithmetic intensity. Indeed the Newton method is based on a repeated construction and solution of the linearization which can lead to both memory bandwidth and communication bottlenecks.

In [252], the FAS nonlinear multigrid method has been successfully applied to the simulation of incompressible flows requiring the solution of the Navier-Stokes equations. The combination of FAS and nonlinear GMRES [261] has been also proposed for the simulation of recirculating flows. The FAS iteration is accelerated by constructing a combination of several previous FAS iterates. In [252], this combination has been successfully applied to the simulation of transient flows. In the future, we want to reinvestigate this question of nonlinear solvers and related convergence acceleration techniques. Limited-memory variants of quasi-Newton methods [54, 182, 190] in combination with multilevel preconditioners will be studied. Applications concern the solution of nonlinear equations in structural mechanics within the PAMSIM project.

5.4.2. Parallelism in time

Motivations

In this section, we consider the numerical solution of evolution problems based on time-dependent partial differential equations. The usual approach for the solution of such transient problems is to use time-stepping methods, consisting in solving a given spatial problem at every time step. However, key problems in Computational Science and Engineering (e.g., in multiphysics simulations) usually involve thousands of time steps, and a scalable parallel solver in space (of either explicit or implicit type) leads unfortunately to unacceptable computational times. Forthcoming extreme scale supercomputers motivate us to reconsider this classical approach. In addition to parallelism in space, the time variable indeed offers a further direction to introduce parallelism. Algorithms that exploit this property fall into the class of time-parallel methods. This class is especially of interest on extreme scale computers for which a large amount of cores will be definitively available. Hence, the design of efficient parallel methods in time is of major concern to tackle the simulation of time-dependent phenomena governed by partial differential equations.

In this section, we consider the case of time-parallel solvers in Computational Fluid Dynamics (CFD) with a specific emphasis on the simulation of compressible flows. As a testbed, we employ Hybrid, a structured-mesh, finite difference code used for direct numerical simulation (DNS) and large eddy simulation (LES) of shock/turbulence interactions developed at the Center for Turbulence Research at Stanford, USA. We want to study the design of efficient time-parallel methods in such a setting. Extensive numerical experiments have been performed to analyse the parallel efficiency of this code on massively parallel platforms; see [32, 33]. We note that the design of time-parallel methods for hyperbolic nonlinear PDEs is still an open question in the community.

Collaboration

This research project is currently being developed with Julien Bodart (ISAE-Supaéro), Serge Gratton and Thibaut Lunet with financial support from Région Midi-Pyrénées, ISAE-Supaéro and CERFACS.

Problem setting

In the class of time-parallel methods, we concentrate on iterative methods to address the solution of time-dependent nonlinear partial differential equations. We refer the reader to the recent comprehensive review paper by Martin Gander [115]. We note that the efficiency of time-parallel methods is usually investigated on simple transient academic problems. Time-parallel methods have recently proved efficient on a few realistic applications (molecular dynamics [53], neutron transport [186] to name a few). However, we are not aware of any analysis or any benchmark results related to the time-parallel simulation of turbulent compressible flows at high Reynolds numbers. Hence, we would like to simultaneously analyse time-parallel methods and propose efficient algorithms in this setting. As discussed during the "Parallel in time methods" workshop in Toulouse (January 11-12 2016), we also aim at proposing a related CFD benchmark problem to be shared within the community to make performance comparisons easier.

Tasks and applications

We briefly present two research directions that we would like to follow. These research directions are generic but applications to CFD are anticipated.

Parareal Multiple shooting methods [115] represent an important class of parallel in time methods. The key idea is to rely on an approximation to the numerical solution on the whole time interval and to update the solution by solving an easier-to-solve approximate system in time. The Parareal method [167] is a popular example related to this class of methods. It aims at computing the numerical solution of general systems of ordinary differential equations written in the form

$$\frac{\partial u}{\partial t} = f(u), \quad t \in [0, T], \quad u(0) = u_0 \tag{5.15}$$

with $u, f : [0,T] \to \mathbb{R}^n$, $u_0 \in \mathbb{R}^n$. The key idea of Parareal is to decompose the time-integration interval $[0,T]$ into N subintervals (not necessarily of equal length) determined by the time-points $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$, to solve an initial value problem on each subinterval concurrently, and to force continuity of the solution branches on successive intervals by means of a Newton procedure. This corresponds to the framework of multiple shooting methods first described in [26]. Given the shooting terms $(U_0, \cdots, U_N)$ which approximate the solution at time $(t_0, \cdots, t_N)$, the N initial value problems can be first solved concurrently

$$\begin{aligned} \frac{\partial u_0}{\partial t} &= f(u_0), \quad u_0(t_0) = U_0, \\ \frac{\partial u_1}{\partial t} &= f(u_1), \quad u_1(t_1) = U_1, \\ &\vdots \\ \frac{\partial u_{N-1}}{\partial t} &= f(u_{N-1}), \quad u_{N-1}(t_{N-1}) = U_{N-1}. \end{aligned}$$

Let denote $u_i(t, U_i)$ the solution at time t on the time interval $[t_{i+1}, t_i]$ with initial condition U_i. To recover the solution of (5.15) we need to impose the matching conditions that lead to a nonlinear system of equations with respect to the shooting variables, i.e.,

$$H(U) = 0, \quad U = (U_0, \cdots, U_N)^T, \tag{5.16}$$

with

$$\begin{aligned} U_0 - u_0 &= 0, \\ U_1 - u_0(t_1, U_0) &= 0, \\ &\vdots \\ U_N - u_{N-1}(t_N, U_{N-1}) &= 0. \end{aligned}$$

Denoting $U^k = (U_0^k, \cdots, U_N^k)^T$ the shooting vector at iteration k, solving the nonlinear system with Newton's method then leads to the following iteration

$$J_H(U^k)(U^{k+1} - U^k) = -H(U^k), \tag{5.17}$$

where $J_H(U^k)$ denotes the Jacobian of H with respect to U^k. It turns out that $J_H(U^k)$ can be easily expressed due to the special structure of H. This finally leads to the form

$$\begin{aligned} U_0^{k+1} &= u_0 \\ U_{n+1}^{k+1} &= u_n(t_{n+1}, U_n^k) + \frac{\partial u_n(t_{n+1}, U_n^k)}{\partial U_n}(U_n^{k+1} - U_n^k), \quad n \in \{0, 1, \cdots, N-1\}. \end{aligned} \tag{5.18}$$

The Parareal method is then recovered if the following two approximations in relation (5.18) are used

$$u_n(t_{n+1}, U_n^k) = F(t_{n+1}, t_n, U_n^k) \tag{5.19}$$

$$\frac{\partial u_n}{\partial U_n}(t_{n+1}, U_n^k)(U_n^{k+1} - U_n^k) \quad = \quad G(t_{n+1}, t_n, U_n^{k+1}) - G(t_{n+1}, t_n, U_n^k), \qquad (5.20)$$

where F and G denote the fine and coarse propagators, respectively ($F(t_b, t_a, U)$ corresponds to a time integration on $[t_a, t_b]$ with U as initial condition at time t_a whereas $G(t_b, t_a, U)$ corresponds to a time integration on $[t_a, t_b]$ with U as initial condition at time t_a with a cheap time-stepping method). This algorithm has received a lot of attention over the past few years. Extensive numerical experiments can be found, e.g., in [186] for neutron transport, in [70, 106, 231] for the Navier-Stokes equations, and in [118] for reservoir simulation. We note that this algorithm can be easily implemented, since it requires only evaluations through the fine and coarse propagators, respectively, leading to a possible black box implementation.

The parallel efficiency of Parareal has been considered theoretically and experimentally. It should be noticed that the parallel efficiency of Parareal on advection dominated academic model problems is found to decrease considerably with an increasing number of processors [231]. Thus improvements of Parareal are required to tackle situations of interest in our setting. The proposed idea is to consider the formalism of the multiple shooting methods (5.16) as a starting point. Instead of using Newton's method, we want to consider methods for unconstrained optimization which are globally convergent (i.e. limited memory variants of Quasi-Newton methods with line search [182] which apply the approximate inverse Jacobian by a sequence of low-rank updates). First attempts to derive a globally convergent algorithm have been proposed in [64]. We want to investigate further in this direction to improve the nonlinear convergence of Parareal and analyse its stability properties.

ParaExp ParaExp [116] belongs to the class of direct time-parallel methods [115] and is only suited to the parallel integration of linear initial-value problems. The key idea of the algorithm relies on the fact that linear initial value problems can be decomposed into homogeneous and nonhomogeneous subproblems. The homogeneous subproblems can be then solved by an exponential integrator, while the nonhomogeneous subproblems can be handled by a classical time-stepping method such as a Runge-Kutta method. We first describe the main ideas of ParaExp and conclude by proposing prospectives. We consider the linear initial value problem

$$\frac{\partial u}{\partial t} = A\ u(t) + g(t), \quad t \in [0, T], \quad u(0) = u_0, \qquad (5.21)$$

where $A \in \mathbb{R}^{n \times n}$ is possibly a large-scale matrix, $u, g : [0, T] \to \mathbb{R}^n$, $u_0 \in \mathbb{R}^n$ and $g(t)$ may be difficult to integrate numerically. Given a time decomposition of $[0, T]$ into N subintervals $[t_0, t_N]$ first the non-overlapping inhomogeneous problems must be solved

$$\frac{\partial v_j}{\partial t} = A\ v_j(t) + g(t), v_j(T_{j-1}) = 0, \quad t \in [t_{j-1}, t_j]. \qquad (5.22)$$

We note that these independent subproblems can be solved in parallel. Then the solution of overlapping homogeneous problems (with $v_0(0) = u_0$)

$$\frac{\partial w_j}{\partial t} = A\ w_j(t), w_j(T_{j-1}) = v_{j-1}(T_{j-1}), \quad t \in [t_{j-1}, t_N] \tag{5.23}$$

are required. The final solution u is then obtained by summation

$$u(t) = v_k(t) + \sum_{j=1}^{k} w_j(t), \quad \text{such that} \quad t \in [t_{k-1}, t_k]. \tag{5.24}$$

The reason why a substantial parallel speedup is possible in ParaExp is that near-optimal approximations of the matrix exponential are known [147]. Hence, the homogeneous problems become very cheap. It is interesting to note that good parallel efficiencies have been reported on both parabolic and hyperbolic problems in [116]. This has been later confirmed in [165], where ParaExp has been extended to address the solution of

$$\frac{\partial u}{\partial t} = A\ u(t) + g(t), \quad t \in [0, T], \quad u(0) = u_0, \tag{5.25}$$

where $A \in \mathbb{R}^{n \times n}$ is possibly a large-scale matrix, $u, g : [0, T] \to \mathbb{R}^{n \times p}$, $u_0 \in \mathbb{R}^{n \times p}$. This situation arises in practice, when considering linear initial-value problems with p multiple initial conditions. The parallel efficiency on hyperbolic problems is especially attractive in our setting, since we know that very few time-parallel methods perform well on such problems.

As a prospective, we aim at extending ParaExp to address the solution of the nonlinear initial value problem

$$\frac{\partial u}{\partial t} = A\ u(t) + B(u(t)) + g(t), \quad t \in [0, T], \quad u(0) = u_0, \tag{5.26}$$

where $A \in \mathbb{R}^{n \times n}$, B, $u, g : [0, T] \to \mathbb{R}^n$, $u_0 \in \mathbb{R}^n$. First attempts have been proposed very recently (see [161] and the talk given by Martin Gander at the Parallel in Time workshop in Toulouse in January 2016), while early related comments have been made in [85, 109].

5.5. Conclusions and outlook

Summary We have presented a few short-term research directions to address, when designing efficient numerical algorithms based on multilevel methods for the solution of partial differential equations on forthcoming architectures. These research lines essentially concerned the improvement of multilevel preconditioners and Krylov subspace solvers *separately*. However, to propose a broader and richer view, we are certainly aware that the numerical solution of partial differential equations requires a tight and close consideration of all the elements of the simulation chain (e.g. modelling, discretization, numerical linear algebra, error analysis and architecture aspects of the computers).

Exploiting these interconnections allows us to explore numerical methods that may be simply impossible to consider by using only staggered approaches. Solver-aware numerical methods and nonlinear preconditioning are two key examples in such a setting. In particular, as emphasized in [171, Chapter 13], it is convenient to consider discretization and preconditioning closely linked together to design efficient numerical methods. We stress that the main advantage of using iterative methods is to stop the iteration process whenever the appropriate accuracy has been reached. We refer the reader to [13, 14, 171] for further enlightening discussions.

High Performance Computing To fully exploit the power of extreme scale machines, we are certainly aware of the current limitations of the proposed research directions in terms of High Performance Computing. Concepts that have been recently introduced to analyse the parallel efficiency of algorithms (roofline performance model [266], execution cache memory model [142]) should be used more systematically; see, e.g., [125, 126]. As pointed out during the DOE ASCR Exascale Mathematics Working Group meeting in 2013[5], new algorithms are also needed. This will require a new mathematical analysis as pointed out in [211]. These algorithms should indeed include asynchronous strategies [112] or stochastic processes, while simultaneously targeting reduced communications. In addition, data locality must be reinforced. We note that recent developments have (re)considered these ideas, e.g., in multigrid methods [4, 5] based on early contributions [39, 40, 69, 144]. Multilevel resiliency [5, 71, 149] is also a forthcoming emerging topic that might be addressed as well.

Stochastic Partial Differential Equations (SPDEs) In this manuscript, we have only considered applications related to deterministic partial differential equations. Partial differential equations with random forcing terms or random coefficients are more and more considered. This can occur in various settings, i.e., when considering either forward problems (e.g. sampling the solution) or inverse problems (e.g. fitting a model to a set of given observations). Elliptic partial differential equations with coefficients given by correlated random fields or reaction-diffusion partial differential equations with random forcing terms are two such examples that are frequently encountered in the field of geosciences. In addition to the mathematical convergence analysis, a key open problem is to propose efficient iterative methods for the solution of such stochastic partial differential equations on massively parallel platforms. Building blocks (Krylov subspace methods, multilevel preconditioning, reuse of subspace information) have been examined in this manuscript but a clever combination of all these elements needs to be developed to address this exciting challenge in the near future.

[5]Report available at http://science.energy.gov/~/media/ascr/pdf/research/am/docs/EMWGreport.pdf

Probabilistic methods Probabilistic representations of solutions of certain partial differential equations are increasingly popular, since they provide powerful analytical tools to establish existence and uniqueness of solutions. By using the Feynman-Kac theorem, the solution of linear parabolic partial differential equations can be obtained by simulating random paths of a stochastic process. This central idea is used in the Probabilistic Domain Decomposition method [3] that has been later extended to handle nonlinear elliptic and nonlinear parabolic problems [1, 2, 209]. Given a partitioning of the computational domain, this method combines a probabilistic approach for evaluating values at the interfaces of the partition and a classical deterministic domain decomposition method; see also the recent contribution [34] for additional algorithmic improvements. Since the simulation of wave propagation phenomena has been covered in the manuscript, we mention the research by Budaev and Bogy [49, 50] on the probabilistic representation of the Helmholtz equation. In addition, recent theory has been developed to handle linear, semi-linear as well as nonlinear partial differential equations by the concept of forward-backward stochastic differential equations, Fourier cosine expansions and wavelets; see [74, 214] and references therein (a backward stochastic partial differential equation is a stochastic differential equation for which a terminal condition, instead of an initial condition, has been specified and its solution consists of a pair of processes). Probabilistic methods indeed offer an important computational advantage, since the algorithms are especially suited for parallel computing. The solution is indeed computed through an expected value over a given finite sample whose elements are independent from each other. This leads to algorithms with extremely low communication overhead, and usually good properties in terms of scalability and fault tolerance. The class of probabilistic methods certainly deserves a specific attention in the near future.

A. Appendix A: Curriculum vitae détaillé

Xavier Vasseur
Né le 8 mars 1971
Nationalité française
Marié, deux enfants

Situation professionnelle
Chercheur
CERFACS
Equipe Algorithmes Parallèles
42 avenue Gaspard Coriolis
F-31057 Toulouse cedex 1
France

E-mail: vasseur@cerfacs.fr

Page web: http://wwww.cerfacs.fr/~vasseur

Formation et diplômes

• Thèse de doctorat en dynamique des fluides et des transferts, Université de Nantes, Nantes, (1998). Mention "Très honorable avec les félicitations du jury". Sujet: Etude numérique de techniques d'accélération de convergence lors de la résolution des équations de Navier-Stokes en formulation découplée ou fortement couplée.
• DEA en "Dynamique des Fluides et Transferts", Université de Nantes, Nantes, (1994). Mention "Très bien".
• Diplôme d'ingénieur, Ecole Centrale de Nantes, Nantes, (1994). Mention "Assez bien".
• Classes préparatoires scientifiques, Lycée Chaptal, Lycée Janson de Sailly, Paris, (1988-1991).
• Baccalauréat, série C, Lycée Jean-Jacques Rousseau, Montmorency (1988). Mention "Assez bien".

Parcours professionnel

• Chef de projet adjoint, Equipe Algorithmes Parallèles, CERFACS, Toulouse, *janvier 2012 - juillet 2015*
• Chercheur sénior, Equipe Algorithmes Parallèles, CERFACS, Toulouse, *depuis mai 2007*

• Post-doctorant, Equipe Algorithmes Parallèles, CERFACS, Toulouse, *mi octobre 2005 - avril 2007*
• Assistant post-doctoral, Ecole Polytechnique Fédérale de Zurich, Suisse, *avril 2002 - août 2005*
Département de Mathématiques, Séminaire de Mathématiques Appliquées
Méthodes de décomposition de domaine pour les éléments finis de type *hp* sur des maillages anisotropes
Responsable: Dr. Andrea Toselli - Bourse du Fonds National Suisse de la Recherche (FNS)
• Assistant post-doctoral, Ecole Polytechnique Fédérale de Lausanne, Suisse, *février 2001 - janvier 2002*
Département de Mathématiques, Chaire de Modélisation et de calcul scientifique
Algorithmes de couplage fluide-structure pour la simulation d'écoulements sanguins au sein d'artères du système cardio-vasculaire humain
Responsable: Prof. Alfio Quarteroni
• Post-doctorant, Université de Paris XI, Orsay, *décembre 1998 - novembre 2000*
Simulations numériques d'écoulements turbulents dans une turbopompe du lanceur Ariane V par la simulation des grandes échelles.
Responsable: Dr. Gérard Albano - Bourse post-doctorale CNES
• Doctorant, Université de Nantes, Nantes, *octobre 1995 - novembre 1998*
Ecole Centrale de Nantes, Laboratoire de mécanique des fluides, Division Modélisation Numérique
Méthodes multigrille pour la résolution des équations de Navier-Stokes
Directeur de thèse: Prof. Jean Piquet - Bourse Docteur Ingénieur du CNRS
• Scientifique du contingent (Marine), Direction des Constructions Navales, Paris, *novembre 1994 - septembre 1995*
Simulations numériques d'écoulements turbulents autour de sous-marins
Encadrant: Dr. Bernard Cardot

A.1. Activités de recherche

A.1.1. Synthèse

Ces dernières années, j'ai pu travailler essentiellement sur les techniques de préconditionnement à base de méthode multigrille et de méthode de décomposition de domaine pour la résolution de systèmes linéaires ou non-linéaires résultant de la discrétisation d'équations aux dérivées partielles. Le but revenait à construire des préconditionnements efficaces et robustes pour des systèmes mal conditionnés issus d'applications réalistes.

Méthodes multigrille

Lors des travaux de doctorat (1995-1998), j'ai pu développer des méthodes multigrille géométriques destinées à la résolution de systèmes linéaires ou non-linéaires issus de

la discrétisation des équations de Navier-Stokes. Les conclusions de ce travail étaient doubles:

- Les méthodes multigrille se sont avérées des préconditionnements robustes pour les méthodes de Krylov lors de la résolution de grands systèmes linéaires [A20,A21][1]. Lors de la résolution de problèmes difficiles (problèmes de convection-diffusion à convection dominante ou problèmes de diffusion pure comportant de fortes anisotropies dans les coefficients), le spectre de la matrice d'itération de la méthode multigrille présentait des valeurs propres isolées élevées qui pénalisent la convergence. Une accélération par méthode de Krylov permettait de capturer en quelques itérations ces valeurs propres et seule cette combinaison conduisait à une vitesse de convergence indépendante du nombre d'inconnues.

- Suivant les résultats obtenus dans le cas linéaire, une combinaison méthode multigrille et méthode à sous-espace a également été proposée lors de la résolution de problèmes non-linéaires (équations de Navier-Stokes). Comme "préconditionnement" de la méthode à sous-espace, un schéma multigrille non-linéaire (Full Approximation Scheme) a été adopté. L'efficacité de cette stratégie a été démontrée sur des cas modèles en mécanique des fluides numérique. Une procédure de résolution couplée des équations de Navier-Stokes a également été proposée [A19].

Une méthode multigrille géométrique utilisée comme préconditionnement pour l'équation aux dérivées partielles dite d'Helmholtz a été également étudiée avec des applications à des problèmes tridimensionnels en sismique [A6,A13]. L'efficacité de cette méthode en termes de passage à l'échelle a notamment été validée dans un environnement massivement parallèle. Enfin, une méthode multigrille algébrique [A1] a été étudiée plus récemment pour la résolution d'une équation aux dérivées partielles elliptique intervenant dans les problèmes de modélisation de réservoirs.

Méthodes de décomposition de domaine

Au cours des années 2002-2005, j'ai pu aborder la thématique des méthodes de décomposition de domaine avec Andrea Toselli, expert de ce sujet. Le but du projet était d'analyser et d'implémenter de nouveaux algorithmes de décomposition de domaine pour la résolution de systèmes provenant de la discrétisation d'équations aux dérivées partielles par des méthodes aux éléments finis de type *hp* sur des maillages étirés. Ces maillages raffinés anisotropiquement sont requis en pratique pour assurer la convergence exponentielle de l'approximation même en présence de singularités et/ou de couches limites.

Les méthodes de décomposition de domaine sont généralement très sensibles aux rapports d'aspect du maillage et leur vitesse de convergence se détériore sévèrement lors d'emploi d'éléments fins par exemple. Un accent particulier a donc été porté sur la

[1] La numérotation fait référence à la liste de publications proposée en Section A.1.2.

robustesse vis à vis des rapports d'aspect du maillage.

Les résultats suivants ont été obtenus:

- Les méthodes de balancing Neumann-Neumann et FETI dans le cadre d'approximations de type *hp* de problèmes scalaires ont été généralisées en dimensions deux et trois [A14,A17,A18]. Des validations détaillées ont démontré leur efficacité et robustesse.

- Une généralisation des méthodes FETI dans le cadre d'approximations de type *hp* de problèmes électromagnétiques bidimensionnels [A16] a été réalisée. Une validation numérique détaillée a démontré l'efficacité et la robustesse de ces méthodes [A15].

Méthodes de Krylov

Mes travaux de recherche plus récents ont trait à l'analyse et au développement de méthodes de Krylov autorisant l'emploi de préconditionnements variables [A10,A12]. Le cas de systèmes linéaires à multiples seconds membres donnés simultanément a également été étudié [A4,A7,A9]. A chaque fois, l'intérêt des nouvelles méthodes a été illustré sur des problèmes concrets en géophysique ou en mécanique des structures, par exemple [A2,A4,A7,A9,S1].

A.1.2. Articles publiés dans des revues internationales à comité de lecture

[*A*1] S. Gratton, P. Hénon, P. Jiránek, X. Vasseur, "Reducing complexity of algebraic multigrid by aggregation", *Numer. Linear Algebra Appl.*. 23, (2016), pp. 501–518.

[*A*2] Y. Diouane, S. Gratton, X. Vasseur, L. N. Vicente, H. Calandra, "A parallel evolution strategy for an Earth imaging problem in geophysics", *Optimization and Engineering*. 17-1, (2016), pp. 3–26.

[*A*3] G. Ramillien, F. Frappart, S. Gratton, X. Vasseur, "Sequential estimation of surface water mass changes from daily satellite gravimetry data", *Journal of Geodesy*, 89-3 , (2015), pp. 259–282.

[*A*4] H. Calandra, S. Gratton, R. Lago, X. Vasseur, L. M. Carvalho, "A deflated minimal residual block method for the solution of non-Hermitian linear systems with multiple right-hand sides", *SIAM J. Sci. Comput.*, 35-5, (2013), pp. S345–S367.

[*A*5] S. Gratton, P. Jiránek, X. Vasseur, "Energy backward error: interpretation in numerical solution of elliptic partial differential equations and convergence of the conjugate gradient method", *Electron. Trans. Numer. Anal.*, 40, (2013), pp 338–355.

[*A*6] H. Calandra, S. Gratton, X. Pinel, X. Vasseur, "An improved two-grid preconditioner for the solution of three-dimensional Helmholtz problems in heterogeneous media", *Numer. Linear Algebra Appl.*, 20, (2013), pp. 663–688.

[*A*7] H. Calandra, S. Gratton, J. Langou, X. Pinel, X. Vasseur, "Flexible variants of block restarted GMRES methods with application to geophysics", *SIAM J. Sci. Comput.*, 34-2 , (2012), pp. A714–A736.

[*A*8] G. Ramillien, L. Seoane, F. Frappart, R. Biancale, S. Gratton, X. Vasseur, S. Bourgogne, "Constrained regional recovery of continental water mass time-variations from GRACE-based geopotential anomalies over South America", *Surveys in Geophysics*, 33-5, (2012), pp. 887–905.

[*A*9] H. Calandra, S. Gratton, R. Lago, X. Pinel, X. Vasseur, "Two-level preconditioned Krylov subspace methods for the solution of three-dimensional heterogeneous Helmholtz problems in seismics", *Numerical Analysis and Applications*, 5, (2012), pp. 175–181.

[*A*10] L. M. Carvalho, S. Gratton, R. Lago, X. Vasseur, "A flexible Generalized Conjugate Residual method with inner orthogonalization and deflated restarting", *SIAM J. Matrix Anal. Appl.*, 32-4, (2011), pp. 1212–1235.

[*A*11] G. Ramillien, R. Biancale, S. Gratton, X. Vasseur, S. Bourgogne, "GRACE-derived surface water mass anomalies by energy integral approach. Application to continental hydrology", *Journal of Geodesy*, 85-6, (2011), pp. 313–328.

[*A*12] L. Giraud, S. Gratton, X. Pinel, X. Vasseur, "Flexible GMRES with deflated restarting", *SIAM J. Sci. Comput.*, 32-4, (2010), pp. 1858–1878.

[*A*13] I. S. Duff, S. Gratton, X. Pinel, X. Vasseur, "Multigrid based preconditioners for the numerical solution of two-dimensional heterogeneous problems in geophysics", *International Journal of Computer Mathematics*, 84-88, (2007), pp. 1167–1181.

[*A*14] A. Toselli, X. Vasseur, "A numerical study on Neumann-Neumann methods for *hp* approximations on geometrically refined boundary layer meshes II: three-dimensional problems", *Mathematical Modelling and Numerical Analysis, M2AN*, 40, 1, (2006), pp. 99-122.

[*A*15] A. Toselli, X. Vasseur, "Robust and efficient FETI domain decomposition algorithms for edge element approximations", *Computation and Mathematics in Electrical Engineering (COMPEL)*, 24, 2, (2005), pp. 396-407.

[*A*16] A. Toselli, X. Vasseur, "Dual-primal FETI algorithms for edge element approximations: two-dimensional h and p finite elements on shape-regular meshes", *SIAM J. Numer. Anal.*, 42, 6, (2005), pp. 2590-2611.

[*A*17] A. Toselli, X. Vasseur, "Domain decomposition preconditioners of Neumann-Neumann type for *hp* approximations on boundary layer meshes in three dimensions", *IMA J. Numer. Anal.*, 24, 1, (2004), pp. 123-156.

[*A*18] A. Toselli, X. Vasseur, "A numerical study on Neumann-Neumann and FETI methods for *hp* approximations on geometrically refined boundary layer meshes in two dimensions", *Comput. Methods Appl. Mech. Engrg.*, 192, 41-42, (2003), pp. 4551-4579.

[*A*19] G.B. Deng, J. Piquet, X. Vasseur, M. Visonneau "A new fully coupled method for computing turbulent flows", *Comput. Fluids*, 30-4, (2001), pp. 445-472.

[*A*20] J. Piquet, X. Vasseur "A non-standard multigrid method with flexible multiple semi-coarsening for the numerical solution of the pressure equation in a Navier-Stokes solver", *Numer. Algorithms*, 24,4, (2000), pp. 333-355.

[*A*21] J. Piquet, X. Vasseur "Multigrid preconditioned Krylov subspace methods for three-dimensional numerical solutions of the incompressible Navier-Stokes equations", *Numer. Algorithms*, 17 (1998) 1-2, pp. 1-32.

A.1.3. Article soumis

[*S*1] S. Gratton, S. Mercier, N. Tardieu, X. Vasseur, "Limited memory preconditioners for the solution of symmetric indefinite problems with application to structural mechanics", CERFACS Technical Report TR/PA/15/48, soumis à *Numer. Linear Algebra Appl.* en juillet 2015, en révision.

A.1.4. Thèse de doctorat

[*T*] X. Vasseur, "Etude numérique de techniques d'accélération de convergence lors de la résolution des équations de Navier-Stokes en formulation découplée ou fortement couplée", Université de Nantes, Ecole Centrale de Nantes, novembre 1998.

A.1.5. Diplôme d'études approfondies

[*D*] X. Vasseur, "Résolution d'une équation d'advection-diffusion par une méthode multigrille (conditions de Dirichlet ou von Neumann sur les frontières) avec estimation de l'erreur de troncature et optimisation mémoire", DEA, Université de Nantes, Ecole Centrale de Nantes, septembre 1994.

A.1.6. Actes de conférences internationales avec comité de lecture

• O. Coulaud, L. Giraud, P. Ramet, X. Vasseur "Deflation and augmentation techniques in Krylov subspace methods for the solution of linear systems", in B.H.V. Topping and P. Ivanyi, (Editor), "Developments in Parallel, Distributed, Grid and Cloud Computing for Engineering", Saxe-Coburg Publications, Stirlingshire, UK, Chapter 11, pp 249-275, 2013.

• A. Mucherino, M. Fuchs, X. Vasseur, S. Gratton, "Variable Neighborhood Search for Robust Optimization and Applications to Aerodynamics", Lecture Notes in Computer Science, Volume 7116, 2012, 8th International Conference, LSSC 2011, Sozopol, Bulgaria, June 6-10, 2011.

• L. Giraud, S. Gratton, X. Pinel, X. Vasseur, "Numerical experiments on a flexible variant of GMRES-DR", PAMM, 7-1 (2007), pp. 1020501-1020502, Sixth International Congress on Industrial Applied Mathematics (ICIAM07).

• A. Toselli, X. Vasseur, "Robust and efficient domain decomposition algorithms for edge element approximations", *11th International IGTE Symposium on Numerical Field Calculation in Electrical Engineering, Seggauberg, Austria, September 13-15 2004*, Department for Fundamentals and Theory in Electrical Engineering, TU Graz, CD-ROM.

• X. Vasseur, "Analysis of a Non-standard Multigrid Preconditioner by Spectral Portrait Computation", *Multigrid Methods, VI, Sixth European Multigrid Conference, Gent, 27-30 September 1999* , "Lecture Notes in Computational Science and Engineering", volume 14, pp. 249-255, E. Dick, K. Riemslagh, J. Vierendeels (eds.), Springer.

• G.B. Deng, J. Piquet, X. Vasseur, M. Visonneau, "A Fully Coupled Procedure with Defect Correction Technique for the Computation of Turbulent Incompressible Viscous Flow past an Airfoil", *Sixth Conference on Numerical Methods for Fluid Dynamics, Oxford, 31 March- 3 Apr. 1998*, pp. 285-291.

• G.B. Deng, J. Piquet, X. Vasseur, M. Visonneau, “Fully Coupled Resolution of the Three-dimensional Navier-Stokes Equations on Cell-centered Colocated Grids by a Nonlinear Multigrid Approach”, *Fifth Annual Conference of the CFD Society of Canada, Victoria, British Columbia, May 1997*, pp. 3-55–3-60.

• X. Vasseur, “A FMG-FAS Procedure for the Fully Coupled Resolution of the Navier-Stokes Equations on Cell-centered Colocated Grids”, *Eighth Copper Mountain Conference on Multigrid Methods, Copper Mountain, Colorado, 6-11 April 1997.*

• J. Piquet, X. Vasseur, “Three-dimensional Multigrid Based Pressure Solver for the Computation of the Flow around the HSVA Tanker“, *Thirteenth GAMM-Seminar "Numerical Treatment of Multi-Scale Problems", Kiel, 24-26 January 1997.* Notes in Numerical Fluid Mechanics, volume 70, pp. 146-155, W. Hackbusch, G. Wittum (eds.), Vieweg.

• J. Piquet, X. Vasseur, “Comparisons between Preconditioned BICGSTAB and a Multigrid Method for the Resolution of the Pressure Equation in a Navier-Stokes Solver”, *Multigrid Methods, V, Fifth European Multigrid Conference, Stuttgart, 1-4 October 1996.* "Lecture Notes in Computational Science and Engineering", volume 3, pp. 225-243, W. Hackbusch, G. Wittum (eds.), Springer.

A.1.7. Actes de conférences nationales avec comité de lecture

• J. Piquet, X. Vasseur , “Résolution de l’équation de pression dans un solveur Navier-Stokes par méthode multigrille utilisée comme solveur ou préconditionneur”, *Treizième Congrès Français de Mécanique, Poitiers, 1-5 septembre 1997*, Tome 3, pp. 119-122.

A.1.8. Rapports techniques

• S. Gratton, S. Mercier, N. Tardieu, X. Vasseur, ”Limited memory preconditioners for the solution of symmetric indefinite problems with application to structural mechanics”, CERFACS Technical Report TR/PA/15/48, 2015.

• Y. Diouane, S. Gratton, X. Vasseur, L. N. Vicente, H. Calandra, ”A Parallel Evolution Strategy for an Earth Imaging Problem in Geophysics”, CERFACS Technical Report TR/PA/15/8, 2015.

• S. Gratton, P. Hénon, P. Jiránek, X. Vasseur ”Reducing complexity of algebraic multigrid by aggregation”, CERFACS Technical Report TR/PA/14/18, 2014.

• O. Coulaud, L. Giraud, P. Ramet, X. Vasseur "Deflation and augmentation techniques in Krylov linear solvers", INRIA Research Report RR-8265, 2013.

• H. Calandra, S. Gratton, R. Lago, X. Vasseur, L. M. Carvalho "A modified block flexible GMRES method with deflation at each iteration for the solution of non-Hermitian linear systems with multiple right-hand sides", CERFACS Technical Report TR/PA/13/15, 2013.

• S. Gratton, P. Jiránek, X. Vasseur "Energy backward error: interpretation in numerical solution of elliptic partial differential equations and behaviour in the conjugate gradient method", CERFACS Technical Report TR/PA/13/16, 2013.

• H. Calandra, S. Gratton, R. Lago, X. Vasseur, and L. M. Carvalho "A deflated minimal block residual method for the solution of non-hermitian linear systems with multiple right-hand sides", CERFACS Technical Report TR/PA/12/45, 2012.

• S. Gratton, P. Jiránek, and X. Vasseur "Energy backward error: interpretation in numerical solution of elliptic partial differential equations and convergence of the conjugate gradient method", CERFACS Technical Report TR/PA/12/3, 2012.

• H. Calandra, S. Gratton, X. Pinel, and X. Vasseur "An improved two-grid preconditioner for the solution of three-dimensional Helmholtz problems in heterogeneous media ", CERFACS Technical Report TR/PA/12/2, 2012.

• H. Calandra, S. Gratton, R. Lago, X. Pinel, and X. Vasseur "Two-level preconditioned Krylov subspace methods for the solution of three-dimensional heterogeneous Helmholtz problems in seismics", CERFACS Technical Report TR/PA/11/80, 2011.

• L. M. Carvalho, S. Gratton, R. Lago, and X. Vasseur "A flexible Generalized Conjugate Residual method with inner orthogonalization and deflated restarting", CERFACS Technical Report TR/PA/11/26, 2011.

• A. Mucherino, M. Fuchs, X. Vasseur, and S. Gratton "Variable neighborhood search for robust optimization and applications to aerodynamics", CERFACS Technical Report TR/PA/11/25, 2011.

• H. Calandra, S. Gratton, J. Langou, X. Pinel, and X. Vasseur "Flexible variants of block restarted GMRES methods with application to geophysics", CERFACS Technical Report TR/PA/11/14, 2011.

• S. Gratton, P. Jiránek, and X. Vasseur "Minimizing the backward error in the energy norm with conjugate gradients", CERFACS Technical Report TR/PA/10/45, 2010.

• L. M. Carvalho, S. Gratton, R. Lago, and X. Vasseur "A flexible Generalized Conjugate Residual method with inner orthogonalization and deflated restarting", CERFACS Technical Report TR/PA/10/10, 2010.

• L. Giraud, S. Gratton, X. Pinel, and X. Vasseur "Flexible GMRES with deflated restarting", CERFACS Technical Report TR/PA/09/111, 2009.

• L. Giraud, S. Gratton, X. Pinel, and X. Vasseur "Flexible GMRES with deflated restarting", CERFACS Technical Report TR/PA/08/128, 2008.

• O. Boiteau, F. Hülsemann, and X. Vasseur "Comparison of the linear algebraic solvers MUMPS and the multifrontal solver of Code ASTER", CERFACS Contract Report TR/PA/06/11, 2006.

• S. Deparis, J.-F. Gerbeau, and X. Vasseur "A Dynamic Preconditioner for Newton-Krylov Algorithms: Application to Fluid-Structure Interaction", INRIA Research Report, RR-5352, 2004.

• A. Toselli and X. Vasseur "Dual-Primal FETI algorithms for edge element approximations: Two-dimensional h and p finite elements on shape-regular meshes", ETHZ, Seminar for Applied Mathematics, Research report 2004-01, 2004.

• A. Toselli and X. Vasseur "A numerical study on Neumann-Neumann methods for hp approximations on geometrically refined boundary layer meshes II: Three-dimensional problems", ETHZ, Seminar for Applied Mathematics, Research report 2003-13, 2003.

• A. Toselli and X. Vasseur "Domain decomposition preconditioners of Neumann Neumann type for hp-approximations on boundary layer meshes in three dimensions", ETHZ, Seminar for Applied Mathematics, Research report 2003-01, 2003.

• A. Toselli and X. Vasseur "A numerical study on Neumann-Neumann and FETI methods for hp-approximations on geometrically refined boundary layer meshes in two dimensions", ETHZ, Seminar for Applied Mathematics, Research report 2002-20, 2002.

• A. Toselli and X. Vasseur "Neumann-Neumann and FETI preconditioners for hp-approximations on geometrically refined boundary layer meshes in two dimensions", ETHZ, Seminar for Applied Mathematics, Research report 2002-15, 2002.

A.2. Activités d'enseignement et d'encadrement doctoral

A.2.1. Enseignement

Le tableau synoptique A.1 résume les activités d'enseignement (C: cours, TD: travaux dirigés et TP: travaux pratiques sous Matlab, Scilab, C ou Fortran 90) pendant les années académiques entre 2006-2007 et 2014-2015. Le volume horaire global par année est également donné.

Année académique	06-07	07-08	08-09	09-10	10-11	11-12	12-13	13-14	14-15
ENSEEIHT/EDP			16	16	16	16	16	16	
ENSEEIHT/PSN							4/4	4/4	4/4
INSA/ANA			13						
ISAE/AMO	25	25	25	25	25	25	12.5	12.5	12.5
ENM/ALG			8	8	8	8	8	8	8
ENM/ANA							18	18	18
ENM/MMC								21/10	21/10
ENM/MPI			5/6	5/6	5/6	5/6	5/6	5/6	
Total C	25	25	30	30	30	30	17.5	37.5	33.5
Total TD								10	10
Total TP			43	30	30	30	52	52	30
Total	**25**	**25**	**73**	**60**	**60**	**60**	**69.5**	**99.5**	**73.5**

Table A.1.: Tableau synoptique des enseignements donnés entre les années universitaires 2006-2007 et 2014-2015 à l'ENSEEIHT (Ecole Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique, d'Hydraulique et des Télécommunications), l'INSA (Institut National des Sciences Appliquées de Toulouse), l'ISAE (Institut Supérieur de l'Aéronautique et de l'Espace, ENSICA) et l'ENM (Ecole Nationale de la Météorologie). EDP: équations aux dérivées partielles (TP, deuxième année, option Mathématiques et Informatique, **chargé de cours**: Serge Gratton), PSN: projet simulation numérique (C/TP, deuxième année, option Mathématiques et Informatique, **chargé de cours**: Xavier Vasseur), ANA: Analyse numérique (TP, première année, **chargé de cours**: Alain Huard), AMO: analyse matricielle et optimisation (C/TD, première année, **chargés de cours**: Serge Gratton et Michel Salaun), ALG: algorithmie (TP, deuxième année, **chargé de cours**: Serge Gratton), MMC: mécanique des milieux continus (C/TD, première année, **chargé de cours**: Xavier Vasseur), MPI: programmation parallèle sur machine à mémoire distribuée (C/TP, deuxième année, option Informatique, **chargé de cours**: Xavier Vasseur). Les chiffres font référence à des heures d'enseignement.

A.2.2. Enseignement à l'étranger

Le tableau synoptique A.2 résume les activités d'enseignement en Suisse (TD: travaux dirigés et TP: travaux pratiques sous Matlab) pendant les années académiques entre 2001-2002 et 2004-2005. Le volume horaire global par année académique est également donné.

Année académique	01-02	02-03	03-04	04-05
EPFL/ANA	48			
EPFL/A I	28			
EPFL/A III	28			
ETHZ/NM I		20		
ETHZ/KA		10	14	14
ETHZ/NM			14	14
ETHZ/LA			28	24
ETHZ/NM			12	26
Total TD	56	30	68	78
Total TP	48			
Total	**104**	**30**	**68**	**78**

Table A.2.: Tableau synoptique des enseignements donnés entre les années universitaires 2001-2002 et 2004-2005 à l'Ecole Polytechnique Fédérale de Lausanne (EPFL) et l'Ecole Polytechnique Fédérale de Zurich (ETHZ). ANA: analyse numérique (TP, **chargés de cours**: Luca Formaggia et Alfio Quarteroni), A I: analyse I (TD, **chargé de cours**: Yves Biollay), A III: analyse III (TD, **chargé de cours**: Yves Biollay), NM I: méthodes numériques I (TD, **chargé de cours**: Rolf Jeltsch), KA: analyse complexe (TD, **chargés de cours**: Pierre Balmer, Daniel Roessler), NM: analyse numérique (TD, **chargés de cours**: Martin Gutknecht, Kasper Nipp, Jörg Waldvogel).

Ecole Polytechnique Fédérale de Lausanne

De mars à juillet 2001, j'ai été assistant pour le cours d'analyse numérique dispensé par Dr. L. Formaggia et Prof. A. Quarteroni pour les étudiants de premier cycle à l'Ecole Polytechnique Fédérale de Lausanne, responsable de la rédaction des exercices, des corrigés des séances théoriques. Je fus également chargé de la rédaction et de la coordination des examens (4 examens lors des sessions de juillet et de septembre 2001). Ce cours vise à apprendre à résoudre pratiquement divers problèmes mathématiques susceptibles de se poser aux ingénieurs. Les thèmes abordés concernent ainsi la résolution de systèmes linéaires et non-linéaires, l'interpolation polynômiale, l'intégration numérique, la résolution d'équations différentielles ordinaires. L'option pédagogique retenue consistait à illustrer systématiquement chaque thème à la fois par des séances d'exercices théoriques et des séances pratiques, où l'étudiant est invité à traiter des problèmes concrets sous Matlab. Les séances hebdomadaires d'exercices pratiques et

théoriques alternaient ainsi chaque semaine et permettaient une meilleure appréhension des questions théoriques. Les cours du semestre d'hiver 2001/2002 concernaient l'analyse pour les problèmes issus des sciences de l'ingénieur (suites, séries, calcul différentiel et intégral de fonctions de plusieurs variables).

Ecole Polytechnique Fédérale de Zürich

Le travail d'assistant à Zürich consiste à encadrer les élèves pendant les séances de cours, à corriger leurs travaux hebdomadaires et de façon mensuelle à assurer deux heures de tutorat pour chaque matière. Les cours enseignés concernent essentiellement les mathématiques pour l'ingénieur (analyse numérique, méthodes numériques et algèbre linéaire) en insistant grâce aux exercices pratiques sous Matlab sur le côté applicatif. Encore une fois, l'illustration numérique semble bénéfique et indispensable pour une majorité des étudiants. Les attentes des étudiants en section Mathématiques ou Physique sont différentes de celles de ceux des filières Génie Civil ou Matériaux. Aussi quand bien même les matières enseignées sont proches ou identiques, il faut savoir s'adapter à son auditorat et aux attentes des étudiants (rappels des bases ou bien compléments théoriques). Aussi je considère cette nouvelle expérience (qui plus est dans une langue étrangère !) comme enrichissante et instructive.

A.2.3. Co-encadrement d'étudiants en master recherche

J'ai pu co-encadrer les étudiants suivants:

- **Oliver Guillet** (étudiant en troisième année, parcours Mathématiques et Informatique à l'Ecole Nationale de la Météorologie, Toulouse, mars - septembre 2015): "Représentation des corrélations d'erreurs d'observation en assimilation de données". Co-encadrement réalisé avec S. Gratton, S. Gürol et A. Weaver.

- **Anne Cassier** (étudiant en troisième année, parcours Mathématiques et Informatique à l'ENSEEIHT, Toulouse, mars - septembre 2014): "Analyse mathématique de méthodes numériques pour la résolution de problèmes issus de l'optimisation et de l'assimilation de données". Co-encadrement réalisé avec S. Gratton et S. Gürol.

- **Aurélien Lecerf** (étudiant en troisième année, parcours Mathématiques et Informatique à l'ENSEEIHT, Toulouse, mars - septembre 2014): "Approche multisimulation pour l'amélioration des performances d'un soveur, analyse d'un algorithme autorisant la parallélisation en temps". Co-encadrement réalisé avec J. Bodart et S. Gratton.

- **Youssef Diouane** (étudiant en troisième année, parcours Mathématiques et Informatique à l'ENSEEIHT, Toulouse, mars - septembre 2011): "Solving a two-dimensional waveform inversion via global optimization methods". Co-encadrement réalisé avec S. Gratton.

- **Mohamed Biari** (étudiant en troisième année, parcours Mathématiques et Informatique à l'ENSEEIHT, Toulouse, mars - septembre 2010): "Résolution de problèmes inverses en géophysique par des méthodes de Quasi-Newton". Co-encadrement réalisé avec S. Gratton.

- **Audrey Bonnement** (étudiant en Mastère II en "Ingénierie Mathématique, Statistique et Economique", Université de Bordeaux I, mars - août 2008): "An iterative method for the null space detection of sparse rank-deficient matrices". Co-encadrement réalisé avec S. Gratton.

- **Antoine Gayou** (étudiant en Mastère II en Mathématiques IMOI-MCS, Université de Pau, mars - août 2007) "Flexible Generalized Conjugate Residual with inner Orthogonalization and Deflated Restarting". Co-encadrement réalisé avec S. Gratton.

A.2.4. Co-encadrement d'étudiants en thèse

- Je co-encadre avec Serge Gratton, Professeur, IRIT, et Julien Bodart enseignant-chercheur à l'ISAE le travail de thèse de **Thibaut Lunet** depuis janvier 2015. Son sujet de thèse est: "Développement de nouvelles stratégies pour le calcul massivement parallèle à l'échelle exa en mécanique des fluides numérique". Date de soutenance prévue: début 2018, Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse, thèse financée par la région Midi-Pyrénées, l'ISAE et le CERFACS, réalisée à l'ISAE et au CERFACS.

- J'ai co-encadré avec Serge Gratton, Professeur, IRIT, et Selime Gürol, chercheuse au CERFACS, le travail de thèse d'**Anne Cassier** de novembre 2014 à juin 2015. Son sujet de thèse était: "Etude mathématique de préconditionnements pour des problèmes de type point-selle de grande dimension avec applications en optimisation et en assimilation de données". Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse, thèse réalisée au CERFACS sur un financement propre CERFACS. Anne Cassier a décidé de se consacrer à la préparation du concours d'aggrégation de mathématiques et a mis un terme à ses travaux de thèse en juin 2015.

- J'ai co-encadré avec Serge Gratton, Professeur, IRIT, et Nicolas Tardieu (EDF) le travail de thèse de **Sylvain Mercier** de novembre 2012 à octobre 2015. Son sujet de thèse était: "Solveurs non-linéaires rapides en mécanique des solides". Date de soutenance: novembre 2015, Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse, thèse CIFRE réalisée au CERFACS et au département Analyses Mécaniques et Acoustique du centre de recherche et de développement d'EDF à Clamart.

- J'ai co-encadré avec Serge Gratton, Professeur, IRIT, le travail de thèse de **Rafael Lago** entre mars 2010 et juin 2013. Son sujet de thèse était: "Méthodes de Krylov avancées pour la résolution de problèmes d'Helmholtz en géophysique."

Date de soutenance: juin 2013, Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse, thèse réalisée au CERFACS et financée par TOTAL. Cette thèse a obtenu le prix Léopold Escande.

- J'ai co-encadré avec Serge Gratton, Professeur, IRIT, le travail de thèse de **Xavier Pinel** entre septembre 2007 et mai 2010. Son sujet de thèse était: "Un préconditionnement à deux niveaux approché pour la résolution de problèmes d'Helmholtz en trois dimensions avec application à la géophysique". Date de soutenance : mai 2010, Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse, thèse réalisée au CERFACS et financée par TOTAL.

A.2.5. Responsabilités pédagogiques

En complément des vacations en écoles d'ingénieurs, je suis également responsable de formations données au CERFACS dans le cadre de la formation interne et externe.

- **Fortran 90** (14 heures). Les nouvelles fonctionnalités du Fortran 90 sont décrites dans ce cours. Ces concepts sont illustrés par des exercices en travaux pratiques. J'ai donné cette formation en 2010, 2011 et 2012 avec Luc Giraud et suis seul responsable depuis 2012.

- **Outils pour la programmation parallèle** (14 heures). La programmation sur machine parallèle à mémoire distribuée grâce à Message Passing Interface (MPI) est détaillée dans ce cours. Les principales caractéristiques de MPI (variables d'envi-ronnement, communications point à point et collectives, types dérivés et topologie des processeurs) sont décrites. Une introduction à Open-MP pour la programmation sur machine à mémoire partagée est également réalisée. Des travaux pratiques sont organisés pour illustrer les notions issues du cours. J'ai donné cette formation en 2010, 2011 et 2012 avec Luc Giraud et suis seul responsable depuis 2012.

- **Résolution de systèmes linéaires (méthodes itératives)** (14 heures). Les méthodes modernes de résolution de systèmes linéaires sont passées en revue (méthodes directes multifrontales et supernodales, méthodes de Krylov et préconditionnement). Cette formation a été donnée avec Luc Giraud en 2010 suite à une demande interne au CERFACS.

- **Introduction au calcul scientifique haute-performance** (8 heures) est un cours organisé pour former les nouveaux arrivants au CERFACS. L'objectif est d'exposer les principes généraux du calcul scientifique et de décrire les principaux outils disponibles. J'ai donné cette formation en 2009, 2010 et 2011 avec Serge Gratton.

J'ai également participé en tant que formateur à des cours en calcul scientifique au niveau national:

- **Algèbre Linéaire Creuse Parallèle** (28 novembre - 02 décembre 2011) (avec E. Agullo, A. Guermouche, P. Ramet, J. Roman, L. Giraud, M. Kern). Ce cours d'une semaine a été donné dans le cadre des sessions de formation de la Maison de la Simulation (40 participants). J'étais impliqué avec Luc Giraud dans la partie consacrée à la résolution des systèmes linéaires (cours et travaux pratiques sur deux journées, 14 heures).

- Un **Cours CERFACS/EDF de Calcul Numérique** a été organisé les 29-30 avril 2009 à Clamart pour EDF. L'idée était d'initier les chercheurs et ingénieurs à l'utilisation de méthodes avancées pour les simulations numériques. Le cours a porté sur l'erreur inverse, les méthodes creuses directes et les méthodes itératives (décomposition de domaine, méthode multigrille). Ce cours de deux jours a été organisé conjointement avec Olivier Boiteau (EDF) et Jean-Philippe Argaud (EDF). Les cours étaient donnés par Philip Avery, Serge Gratton et moi-même.

A.3. Participation à la vie scientifique et responsabilités collectives

A.3.1. Diffusion de connaissances et animation scientifique

Workshops

• **Parallel in time methods**, Janvier 2016. Ce workshop fait partie du semestre thématique CIMI intitulé "High performance linear and nonlinear methods for large scale applications" se déroulant entre juin 2015 et janvier 2016. Je suis co-organisateur de ce workshop avec C. Besse (IMT, Toulouse) et S. Gratton.

• **Sparse Days** (2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014), CERFACS, Toulouse. Membre du comité local d'organisation (environ 50 participants chaque année). Ce workshop sur un format de deux jours rassemble des orateurs internationaux autour de sujets proches de l'algèbre linéaire, optimisation numérique et du calcul scientifique intensif.

• **Recent advances on Optimization 2013** (July 24-26th 2013). Cette conférence internationale d'une durée de trois jours a rassemblé 120 participants autour de l'optimisation numérique. Cette conférence a été partiellement financée par la fondation RTRA STAE. J'étais membre du comité local d'organisation.

• **15th Austrian-French-German conference on Optimization**, Toulouse, 19-23 septembre 2011. J'étais en charge des relations avec la mairie de Toulouse pour l'organisation de la cérémonie de bienvenue (140 participants).

• **Workshop final du projet ANR Solstice** en juin 2010, CERFACS, Toulouse. Membre du comité local d'organisation (40 participants).

• **Workshop RTRA STAE** "Méthodes avancées et perspectives en optimisation non-linéaire et contrôle" 3-5 février 2010, Toulouse. Membre du comité local d'organisation en collaboration avec la fondation RTRA STAE (120 participants). Serge Gratton et moi-même avions reçu une bourse de la fondation RTRA STAE pour organiser ce workshop international. 15 orateurs internationaux dans le domaine de l'optimisation ont été invités à cette occasion.

Exposés invités

• X. Vasseur, "Combination of multilevel methods and Krylov subspace methods for acoustic full waveform inversion in seismics", WAVES 2015, Karlsruhe, Germany, July 20-24th 2015. En collaboration avec H. Calandra, Y. Diouane, S. Gratton.

• X. Vasseur, "Parallel solution of linear initial-value problems with the ParaExp algorithm", MATHIAS, Paris, October 29-31 2014. En collaboration avec J. Bodart, S. Gratton, A. Lecerf.

• X. Vasseur, "Geometric and algebraic multilevel methods for problems in seismic and reservoir modelling", MATHIAS, Paris, October 23-25 2013. En collaboration avec H. Calandra, S. Gratton, P. Hénon and P. Jiránek.

• X. Vasseur, "Development of mathematical models for Exascale and beyond", Examath DOE workshop, Washington, USA, August 21-22 2013. En collaboration avec I. Duff, S. Gratton and D. Titley-Péloquin.

• X. Vasseur, "Massively parallel computations for the solution of three-dimensional Helmholtz heterogeneous problems in seismic imaging", MATHIAS, Paris, October 26-28 2012. En collaboration avec H. Calandra, S. Gratton, R. Lago and X. Pinel.

• X. Vasseur, "Multilevel preconditioned Krylov subspace methods for the solution of three-dimensional heterogeneous Helmholtz problems in seismics", First Russian-French Conference on Mathematical Geophysics, Mathematical Modeling in Continuum Mechanics and Inverse Problems, June 18-22th 2012. En collaboration avec H. Calandra, S. Gratton, R. Lago and X. Pinel.

• X. Vasseur, "An approximate two-level preconditioner combined with flexible Krylov subspace methods for the solution of heterogeneous Helmholtz problems on massively parallel computers", ESF OPTPDE Workshop, Wuerzburg, Germany, September 26-28th 2011. En collaboration avec H. Calandra, S. Gratton and X. Pinel.

• X. Vasseur, "Architectures massivement parallèles: quelques questions algorithmiques ouvertes", Ateliers de Modélisation de l'Atmosphère (AMA) 2011, Toulouse, February 8th 2011. En collaboration avec L. Giraud.

• X. Vasseur, "Null space computation of sparse singular matrices with MUMPS", MUMPS User Group Meeting, Toulouse, April 16th 2010. En collaboration avec P. Amestoy, S. Gratton and J-Y. L'Excellent.

• X. Vasseur, "Massively parallel computations for the solution of Helmholtz problems in geophysics", *Department of Mathematics, TU Delft, April 2009.* En collaboration avec H. Calandra, S. Gratton and X. Pinel.

• X. Vasseur, "A two-grid method used as a preconditioner for the solution of Helmholtz problems", *Numerical Analysis Seminar, CWI Amsterdam, April 2009.* En collaboration avec H. Calandra, S. Gratton and X. Pinel.

• X. Vasseur, "Multigrid preconditioned Krylov subspace methods for the numerical solution of the Helmholtz equation in geophysics", *Seminar für Analysis und Numerik, Department of Mathematics, University of Basel, June 2007.* En collaboration avec H. Calandra, I.S. Duff, S. Gratton and X. Pinel.

• X. Vasseur, "Subspace acceleration for linear and nonlinear multigrid methods", *Numerical Analysis seminar, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne, November 2000.*

• X. Vasseur, "Quelques applications de la méthode multigrille en mécanique des fluides numérique", *Institut Français du Pétrole, Rueil-Malmaison, May 27th 1999.*

Séminaires

• 2007-2010: organisation des séminaires internes de l'équipe Algorithmes Parallèles.

A.3.2. Fonctions d'intérêt collectif

• Depuis 2015: membre du "Comité d'Entreprise" au sein du CERFACS.
• Depuis 2014: représentant du CERFACS au sein du consortium MUMPS.
• 2009-2014: membre du "Comité d'Evaluation des chercheurs et ingénieurs du CERFACS".
• Depuis 2009: membre du laboratoire commun avec l'IRIT.
• 2008: membre du laboratoire commun avec l'INRIA Bordeaux Sud-Ouest.

A.3.3. Expertise

Reviewer pour les journaux d'analyse numérique, d'algèbre linéaire et de calcul scientifique suivants

• ACM Transactions on Mathematical Software (2015)
• Applied Numerical Mathematics (2009, 2010)

• BIT (2010)
• Computation and Visualization in Science (2006)
• Concurrency and Computation: Practice and Experience (2010)
• European Journal of Mechanics B (2008)
• Journal of Computational and Applied Mathematics (2010)
• Linear Algebra and Applications (2009)
• Numerical Algorithms (2013, 2014)
• Numerical Linear Algebra with Applications (2012, 2013)
• Parallel Computing (2006, 2007, 2009, 2010, 2011, 2012)
• SIAM Journal on Numerical Analysis (2011)
• SIAM Journal on Scientific Computing (2004, 2008, 2013, 2014).

Reviewer pour les conférences suivantes

• CSE08: IEEE 11th International Conference on Computational Science and Engineering
• Euro-Par15, membre du comité scientifique sur le thème "Numerical methods and applications"
• Euro-Par12
• Euro-Par09
• Supercomputing 2012
• Supercomputing 2011
• VECPAR16: 12th International Meeting High Performance Computing for Computational Science
• VECPAR14: 11th International Meeting High Performance Computing for Computational Science
• VECPAR12: 10th International Meeting High Performance Computing for Computational Science
• VECPAR10: 9th International Meeting High Performance Computing for Computational Science
• VECPAR08: 8th International Meeting High Performance Computing for Computational Science.

Expert auprès de l'Agence Nationale de la Recherche

J'ai pu expertiser deux projets auprès de l'Agence Nationale de la Recherche au sein du programme COSINUS (année 2010) et un autre au sein du programme "Défi Société de l'information et de la communication - Axe Données massives et calcul intensif": enjeux et synergies pour la simulation numérique (année 2015).

Participations aux jurys de thèse suivants

• **Sylvain Mercier**, "Fast nonlinear solvers in solid mechanics". Université Paul Sabatier, Toulouse. Thèse soutenue en novembre 2015. Co-directeur de thèse. Le directeur de thèse était Prof. Serge Gratton.

• **Youssef Diouane**, "Globally convergent evolution strategies with application to an Earth imaging problem in geophysics". Institut National Polytechnique de Toulouse. Thèse soutenue en octobre 2014. Membre du jury. Le directeur de thèse était Prof. Serge Gratton.

• **Pablo Salas**, "Physical and numerical aspects of thermoacoustic instabilities in annular combustion chambers". Université de Bordeaux I. Thèse soutenue en novembre 2013. Membre du jury. Le directeur de thèse était Prof. Luc Giraud.

• **Selime Gürol**, "Solving regularized nonlinear least-squares problems in dual space with application to variational data assimilation". Institut National Polytechnique de Toulouse. Thèse soutenue en juin 2013. Membre du jury. Le directeur de thèse était Prof. Serge Gratton.

• **Rafael Lago**, "A study on block flexible iterative solvers with application to Earth imaging problem in geophysics". Institut National Polytechnique de Toulouse. Thèse soutenue en juin 2013. Co-directeur de thèse. Le directeur de thèse était Prof. Serge Gratton.

• **Sylvie Detournay**, "Multigrid methods for zero-sum two player stochastic games". Ecole Polytechnique (Palaiseau), INRIA Saclay. Thèse soutenue en septembre 2012. Membre du jury. Le directeur de thèse était Prof. Marianne Akian.

• **Mathieu Chanaud**, "Conception d'un solveur haute performance de systèmes linéaires creux couplant des méthodes multigrilles et directes pour la résolution des équations de Maxwell 3D en régime harmonique discrétisées par éléments finis." Université de Bordeaux I, Département d'informatique, thèse soutenue en octobre 2011. Membre du jury. Le directeur de thèse était Prof. Jean Roman.

• **Mikko Byckling**, "Preconditioning for Standard and Two-Sided Krylov Subspace Methods", Helsinki University of Technology, Institute of Mathematics. Thèse soutenue en janvier 2011. J'étais un des deux rapporteurs de la thèse. Le directeur de thèse était Prof. Marko Huhtanen.

• **Xavier Pinel**, "A perturbed two-level preconditioner for the solution of three-dimensional heterogeneous Helmholtz problems with applications to geophysics". Thèse soutenue en mai 2010, Institut National Polytechnique de Toulouse. J'étais co-encadrant et le directeur de thèse était Prof. Serge Gratton.

• **Mélodie Mouffe**, "Multilevel optimization in infinity norm and associated stopping criteria". Thèse soutenue en février 2009, Institut National Polytechnique de Toulouse. Membre du jury. Le directeur de thèse était Prof. Serge Gratton.

B. Appendix B: five selected papers

B.1. A new fully coupled method for computing turbulent flows

PERGAMON

Computers & Fluids 30 (2001) 445–472

computers & fluids

www.elsevier.com/locate/compfluid

A new fully coupled method for computing turbulent flows

G.B. Deng [a], J. Piquet [a,*], X. Vasseur [b,1], M. Visonneau [a]

[a] *Laboratoire de Mécanique des Fluides de l'Ecole Centrale de Nantes UMR 6598 1, rue de la Noë, BP 92101 F-44321 Nantes Cedex 3, France*

[b] *Département de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

Received 8 September 1999; received in revised form 1 March 2000; accepted 25 April 2000

Abstract

This work discusses the computation of the steady, turbulent, two-dimensional incompressible viscous flow on structured cell-centered collocated grids. A rather new computational approach – the so-called fully coupled procedure with defect correction technique – is presented as an alternative both to classical decoupled approaches (SIMPLE [Int. J. Heat Mass Transfer 15 (1972) 1787–1806], PISO [J. Comput. Phys. 62(1) (1986) 40–65] and variants) and to weakly coupled solution methods of Vanka type [Fifth symposium on Turbulent Shear flows, 1985. p. 20:27–32, J. Comput. Phys. 65 (1986) 138–158]. Its main features will be highlighted and detailed. This strategy is evaluated on two demanding test cases (the simulation of the separated flow past an AS240-B airfoil at high incidence (19°, $Re = 2 \times 10^6$) and the simulation of the wake flow behind a two-dimensional hill ($Re = 60\,000$)), for which documented experimental data are available. Both robustness and computational efficiency of the new approach are shown. © 2001 Elsevier Science Ltd. All rights reserved.

1. Introduction

Due to unceasing and fast advances in computer speed and available memory, the numerical solution of the Navier–Stokes equations for an incompressible viscous fluid by fully coupled (FC) procedures tends to be an attractive and emerging trend in computational fluid dynamics (CFD). Concisely, their main advantage consists in an increased robustness due to an implicit and global treatment of the pressure–velocity coupling. Therefore, this framework may represent a promising algorithmic alternative to standard decoupled (DC) approaches (SIMPLE [31], SIMPLER [30],

* Corresponding author. Tel.: +33-2-40-37-16-33; fax: +33-2-40-37-25-23.

E-mail addresses: jean.piquet@ec-nantes.fr (J. Piquet), xavier.vasseur@epfl.ch (X. Vasseur).

[1] Also Corresponding author. Tel.: +41-21-693-5509; fax: +41-21-693-4303.

0045-7930/01/$ - see front matter © 2001 Elsevier Science Ltd. All rights reserved.
PII: S0045-7930(00)00025-6

PISO [21] and their derivatives, also called pressure-correction or distributive-type iteration methods) and to weakly coupled methods (box or box-line relaxation schemes, initially proposed by Vanka [46,47] giving rise to numerous related works, see among others, Refs. [8,15,28,44,45]). Thus, this approach deserves to be analyzed on realistic configurations to estimate its true potential. In the following, the simulation of turbulent viscous flow past moderately complex two-dimensional geometries will be considered.

Truncated Newton methods or inexact Newton–Krylov methods are probably the most popular examples of FC procedures [7,23,26,32]. Nowadays, these approaches, whose linear and non-linear convergences are considered as quite robust (see Ref. [5] for the mathematical aspects), consist in a keystone of major research and industrial codes devoted either to large simulations of coupled phenomena in CFD or more generally to scientific computing. Thus, the scalable non-linear equation solver (SNES) package, a subset of the C library portable, extensible toolkit for scientific computing (PETSc) (see http://www-fp.mcs.anl.gov/petsc/) proposes such an approach with various Krylov solvers and preconditioners as a means to design portable, parallel efficient solution strategies for solving large-scale non-linear problems. Also, an important CFD research project held at Sandia National Laboratories – the simulation of complex three-dimensional fluid flow, heat and mass transfer with complex bulk fluid and surface chemical reaction kinetics with the massively parallel MPSalsa code [39] on Sandia's Teraflop computer (ASCI Red Intel's supercomputer with 9152 processors) – includes this choice as a ground solver inside the Aztec iterative linear algebra library [20]. Very large-scale problems have been successfully treated with such a solution strategy [14,40]. However, the main admitted numerical bottleneck consists in limiting the computational effort devoted to the numerical formation of the jacobian and in designing a robust and reliable preconditioner for the jacobian.

In the class of FC methods, few alternatives to Newton-like or Newton–Krylov methods have been proposed over the past few years. Following the idea initially proposed by Schneider and Zedan [38], Karki and Mongia [22] built an FC system with a true equation for the pressure variable deduced from the continuity equation. This original approach – combined with a direct solver – has been evaluated on various two-dimensional configurations (lid-driven cavity and skewed planar channel). This study has shown that the FC method was efficient on smoothly non-orthogonal grids but lost its robustness on highly stretched and non-orthogonal grids. A second example was the approach proposed by Hanby et al. [19] for studying laminar flows in the cavities between the rotating discs in gas turbine engines. An FC system – retaining the continuity equation – is built and evaluated on various geometries (rotating inner cylinder and rotating cavity problems). The results obtained have shown both the computational efficiency of FC methods for simulating flows at moderate Reynolds numbers and a pronounced deterioration in performances for higher Reynolds numbers, requiring a more effective preconditioner for the coupled linear system.

In the early nineties, an FC method with iterative linear solvers has been proposed by Deng et al. [9]. While the simulation of turbulent viscous flows on refined meshes by classical DC methods (SIMPLE, PISO, etc.) led to non-linear convergence stagnations, the FC method produced a very efficient way to compute steady flows, as proved by the simulation of the turbulent axisymmetric flow around an Afterbody 3 hull [13]. This last work is of relevant interest since it is, to our knowledge, the first application of FC methods for simulating turbulent viscous flows on moderately complex geometries by the resolution of the incompressible Reynolds-averaged Navier–

Stokes equations. For this reason, the FC strategy is retained in the following. However, Deng et al. have been faced with major drawbacks during further developments. The simulation of turbulent viscous flows on refined, highly stretched and curvilinear grids led to dramatic failures in the solution phase, mainly due to the high condition number of the FC system and its high non-normality. This difficulty was circumvented by using dual time stepping in order to enhance the diagonal dominance of the considered linear system. The price to pay was a strong increase in computational time, greatly penalizing the interest of the approach and making it unsuitable for the investigation of unsteady flows. For these reasons, the method was abandoned for a while. The aim of this paper is thus to present appropriate cures to increase the robustness of this strategy for computing steady turbulent viscous flow on two-dimensional configurations. The FC method has also been enhanced for unsteady applications: this subject will be detailed in a forthcoming paper.

The FC method originally proposed in Ref. [9] will be presented and summarized in Part 2. New developments concerning both the implementation of second-order accurate discretization schemes and the solution strategy for the resulting linear system are discussed and outlined, respectively, in Parts 3 and 4. The various numerical experiments will be presented in Part 5. Performances of the FC method will be therein compared with respect to those of classical DC methods on two demanding test cases: the simulation of the separated flow past an airfoil at high incidence (19°, $\mathrm{Re} = 2 \times 10^6$) and the simulation of the wake flow behind an hill ($\mathrm{Re} = 60\,000$). Finally, concluding remarks and perspectives will be drawn in Part 6.

2. The fully coupled method

In this section, the numerical background of this study is presented. This quick overview will lead to a set of dimensionless discrete equations. Then, the FC formulation already presented in Ref. [9] will be explained. The construction of the FC system will be more specifically described.

2.1. Numerical background

2.1.1. Governing equations

The steady Reynolds-averaged Navier–Stokes equations for an incompressible viscous fluid are considered. The master equations can be written in a convective dimensionless form (with summation over repeated indices):

$$\frac{\partial u_i}{\partial x_i} = 0, \tag{1}$$

$$\left(u_j - \frac{\partial v_\mathrm{T}}{\partial x_j}\right)\frac{\partial u_i}{\partial x_j} = \frac{1}{R_\mathrm{eff}}\frac{\partial^2 u_i}{\partial^2 x_j} + \frac{\partial v_\mathrm{T}}{\partial x_j}\frac{\partial u_j}{\partial x_i} - \frac{\partial}{\partial x_i}P, \tag{2}$$

where

$$\frac{1}{R_\mathrm{eff}} = \frac{1}{\mathrm{Re}} + v_\mathrm{T}, \qquad P = \frac{p}{\rho} + \frac{2}{3}K. \tag{3}$$

Eqs. (1) and (2) involve the fluid density ρ, the mean Cartesian velocity components u_i, the mean pressure p, the Reynolds number Re, the effective Reynolds number R_{eff}. Eq. (2) has been obtained assuming a linear eddy-viscosity hypothesis for the turbulent Reynolds stresses:

$$-\overline{u_i' u_j'} = \nu_T \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) - \frac{2}{3} K \delta_{ij}, \tag{4}$$

where δ_{ij} represents the Kronecker symbol, K, the turbulent kinetic energy ($K = \overline{u_i' u_i'}/2$) and ν_T, the turbulent eddy viscosity which defines the eddy-viscosity model to be used. As shown by experimental studies, the configurations presented in Part 5 involve strong adverse pressure gradients. Therefore, K–ω model of Wilcox [53] and the shear stress transport (SST) variant of Menter [27] have been adopted here for both their ability to simulate such phenomena and their numerical robustness.

2.1.2. Partial coordinate transformation

The task of performing numerical simulations past moderately complex geometries generally requires a coordinate transformation which makes the application of boundary conditions easier. This transformation maps the physical domain, in which the flow is studied, into a rectangular one. While dependent variables (Cartesian components of velocity and pressure) are left unchanged, independent variables are taken as the curvilinear, body-fitted coordinates noted $\xi^i = (\xi^1, \xi^2) = (\xi, \eta)$. Byproducts of this transformation are the normalized contravariant basis and the contravariant metric tensor defined respectively, by the vectors $\overrightarrow{g^i}$, such as $J\overrightarrow{g^i} = \overrightarrow{b^i}$ and by $g^{ij} = \overrightarrow{g^i} \cdot \overrightarrow{g^j}$, where J is the jacobian of the transformation from the computational space of coordinates ξ^i to the physical space of coordinates $\vec{x} = (x_1, x_2)$ and $\overrightarrow{b^i}$ area vectors ((i, j, k) in cyclic order) such as

$$\overrightarrow{b^i} = \frac{\partial \vec{x}}{\partial \xi^j} \times \frac{\partial \vec{x}}{\partial \xi^k}, \qquad J = \frac{\partial \vec{x}}{\partial \xi^i} \cdot \left\{ \frac{\partial \vec{x}}{\partial \xi^j} \times \frac{\partial \vec{x}}{\partial \xi^k} \right\}. \tag{5}$$

This coordinate transformation induces in the computational space the following momentum equation for the generic ϕ variable:

$$g^{ii} \frac{\partial^2 \phi}{\partial^2 \xi^i} = \left[R_{\text{eff}} \frac{b_j^k}{J} \left(u_j - \frac{\partial \nu_T}{\partial x_j} \right) - \frac{1}{J} \frac{\partial}{\partial \xi^i} (J g^{ik}) \right] \frac{\partial \phi}{\partial \xi^k} + \underline{R_{\text{eff}} \frac{\partial \phi}{\partial t}} + S_\phi - g_{i \neq k}^{ik} \frac{\partial^2 \phi}{\partial \xi^i \partial \xi^k}, \tag{6}$$

where S_ϕ represents a source term. Note that although only steady applications will be considered, a dimensionless time-dependent term is introduced in Eq. (6) (underlined term). This aspect is mainly motivated for enhancing numerical properties and will be outlined in the forthcoming parts.

2.1.3. Grid lay-out and discrete equations

A monoblock structured grid is used for both applications. Finite difference discretization schemes are considered here. Unknowns are collocated and cell-centered: all variables are stored at each center of control volume. Fig. 1 presents this layout with some helpful notations. Neighboring points of each cell-center C that are involved by the discretization processes are

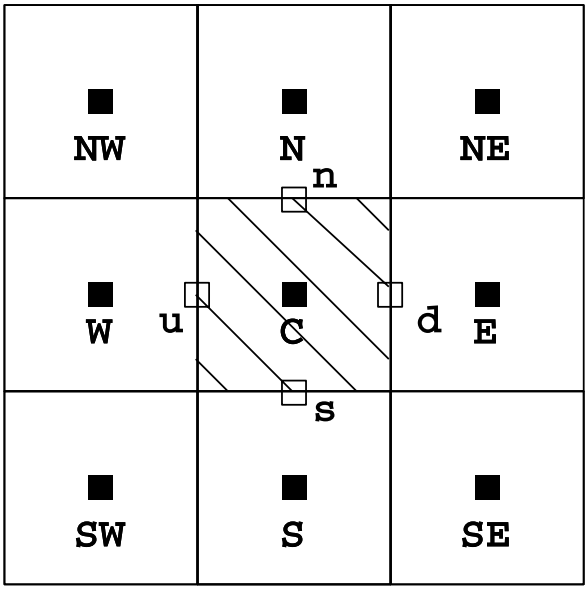


Fig. 1. Control volume of center C and flux reconstruction at the interfaces noted (d, n, u, s).

denoted by (nb). With this convention, the final discretization form of the momentum equation (6) reads as follows (assuming a Picard linearization):

$$\phi_C = \sum_{nb} K_{nb}^{\phi} \phi_{nb} - K_C^{\phi} \left[S_{\phi} - e_1^{\phi} \phi_C^{n-1} \right] \tag{7}$$

with

$$K_{nb}^{\phi} = \frac{C_{nb}^{\phi}}{(1 + e_1^{\phi} C_C^{\phi})}, \qquad e_1^{\phi} = \frac{R_{\text{eff}}}{\tau_{\text{f}}^{\phi}}, \tag{8}$$

where τ_{f}^{ϕ} denotes the false time-step, $(n-1)$, the previous non-linear iteration, each non-indexed variable being evaluated at the current non-linear iteration (n), C_{nb}^{ϕ}, the influence coefficient at the (nb) point for the variable ϕ and S_{ϕ}, the source term gathering contributions of pressure and turbulence variables. This normalized discretization form (7) will serve as a starting point in the following. The discretization of the equations for the turbulence variables follows the same practice and will be omitted for the sake of brevity. For more details concerning the numerical background involved here, the reader is referred to Ref. [12]. The next step consists in detailing the adopted numerical method: the FC formulation.

2.2. Fully coupled formulation

2.2.1. The projection form

The starting point of the formation of the FC system is just a rewriting of the discrete momentum equation (7) with U_i the discrete Cartesian velocity component instead of the discrete variable ϕ. It leads to the following projection form (note that the pressure gradient has been now explicitly expressed):

$$U_{iC} = \widehat{U_{iC}} - K_C^{U_i} R_{\text{eff}} \frac{\partial P}{\partial x_i}, \tag{9}$$

$$\widehat{U_{iC}} = \sum_{nb} K_{nb}^{U_i} U_{inb} + K_C^{U_i} \left[e_1^{U_i} U_{iC}^{n-1} - S_{U_i} \right]. \tag{10}$$

The pseudo-velocity field $\widehat{U_i}$ gathers all contributions of the right-hand side of the momentum equations, except the pressure gradient contribution. Eqs. (9) and (10) can be recast under the symbolic form:

$$\mathbf{U} = \widehat{\mathbf{U}} - \mathbf{K_p G P} \quad \text{with } \mathbf{U} = (\mathbf{U}, \mathbf{V}) \quad \mathbf{K_p} = \mathbf{R}_{\text{eff}}(\mathbf{K_C^U}, \mathbf{K_C^V}), \tag{11}$$

$$\widehat{\mathbf{U}} = \mathbf{C U} + \mathbf{S}_{\widehat{\mathbf{U}}} \quad \text{with } \widehat{\mathbf{U}} = (\widehat{\mathbf{U}}, \widehat{\mathbf{V}}) \quad \mathbf{S}_{\widehat{\mathbf{U}}} = (\mathbf{S}_{\widehat{\mathbf{U}}}, \mathbf{S}_{\widehat{\mathbf{V}}}) \tag{12}$$

with

$$\mathbf{S}_{\widehat{\mathbf{U}}} = \mathbf{K_C^U} \left[e_1^U U_C^{n-1} - S_U \right], \qquad \mathbf{S}_{\widehat{\mathbf{V}}} = \mathbf{K_C^V} \left[e_1^V V_C^{n-1} - S_V \right], \tag{13}$$

where the $\mathbf{C}$ operator gathers the off-diagonal influence coefficients ($\mathbf{K_{nb}^U}$ and $\mathbf{K_{nb}^V}$) of the convection-diffusion scheme and $\mathbf{G}$ denotes the gradient operator.

2.2.2. The pressure equation

The continuity equation (1) requires a flux specification at the interfaces of every control volume (Fig. 1). To build these fluxes, neighboring cell centered quantities must be used. When a linear interpolation between two cell centers is used, a discrete set of equations is obtained with no dependencies between adjacent pressure–velocity variables. One direct consequence is the possibility to obtain non-physical solutions. A cure to solve this odd–even decoupling problem is to follow the Rhie and Chow practice [33]. In this formulation, each flux at the generic interface int (d, u, n or s) is written under the following form (where $\mathscr{U}^1$ and $\mathscr{U}^2$ are the discrete contravariant velocity):

$$\mathscr{U}^i_{\text{int}} = \widehat{\mathscr{U}^i}_{\text{int}} - \left[K_C^{U_i} R_{\text{eff}} \left(b_1^i \frac{\partial P}{\partial x} + b_2^i \frac{\partial P}{\partial y} \right) \right]_{\text{int}}, \tag{14}$$

where each pseudo-flux ($\widehat{\mathscr{U}^1}$, $\widehat{\mathscr{U}^2}$) is linearly interpolated with the help of pseudo-velocity variables known at each cell center close to the considered interface. The construction of $K_C^{U_i}$ and R_{eff} coefficients is similar. However each pressure gradient is discretized at each interface by central differencing. This mixed mechanism (interpolation–discretization) thus involves a pseudo-physical construction. Nevertheless, it is adopted as a common practice. After discretization of the pressure gradient, the final form of the discrete continuity equation can be written in the compact form:

$$\frac{\partial}{\partial \xi^i} \left[b_j^i \widehat{U_j} \right] - \frac{\partial}{\partial \xi^i} \left[a^{ij} \frac{\partial P}{\partial \xi^j} \right] = 0 \quad \text{with } a^{ij} = J g^{ij} K_C^{U_i} R_{\text{eff}}. \tag{15}$$

A discrete form reads as follows:

$$\sum_{nb} C_{nb}^{P\widehat{U}} \widehat{U}_{nb} + \sum_{nb} C_{nb}^{P\widehat{V}} \widehat{V}_{nb} + \sum_{nb} K_{nb}^{PP} P_{nb} = S_P, \qquad S_P = \frac{\partial}{\partial \xi^i} \left[a^{ij} \frac{\partial P}{\partial \xi^j} \right]_{i \neq j}, \tag{16}$$

where the K_{nb}^{PP} coefficients denote the influence coefficients resulting from the discretization of the generalized Laplace pressure operator, C_{nb}^{PU}, C_{nb}^{PV} the influence coefficients corresponding to $\widehat{U}$ and $\widehat{V}$ variables and S_P the cross-derivative pressure terms. Finally, the following algebraic form is deduced where **D** denotes the divergence operator :

$$\mathbf{D}\widehat{\mathbf{U}} - \mathbf{D}\mathbf{K}_\mathbf{p}\mathbf{G}\mathbf{P} = \mathbf{S}_\mathbf{P}. \tag{17}$$

2.2.3. The fully coupled system

It is now possible to build the algebraic form of the FC system, gathering the projection form equations (11), the definition of the pseudo-velocity variables (12) and the pressure equation (17):

$$\begin{bmatrix} \mathbf{I} & -\mathbf{C} & \mathbf{O} \\ -\mathbf{I} & \mathbf{I} & \mathbf{K}_\mathbf{p}\mathbf{G} \\ \mathbf{D} & \mathbf{O} & -\mathbf{D}\mathbf{K}_\mathbf{p}\mathbf{G} \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{U}} \\ \mathbf{U} \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{\widehat{\mathbf{U}}} \\ \mathbf{O} \\ \mathbf{S}_\mathbf{P} \end{bmatrix}. \tag{18}$$

The major difficulty of the incompressible Navier–Stokes equations consists in the absence of any pressure term in the continuity equation. This induces in the linearized system a zero pressure diagonal block. The resulting linear system is thus not an M matrix [48], making the use of standard iterative linear solver difficult. A true equation (17) is therefore deduced to avoid this drawback. As a consequence, this choice introduces the pseudo-velocities as truly implicit variables. The price to pay is thus an increase in the size of the linear system.

Avoiding the introduction of pseudo-velocities in the Rhie and Chow interpolation would yield a wider set of neighboring points involved in the flux reconstruction step. Enhancement of this formulation requires a new flux reconstruction mode: this idea has been developed by Deng et al. [10,11], leading to the so-called consistent physical interpolation (CPI) method. A further study should be undertaken to compare the advantages and drawbacks of both formulations.

3. Defect correction

In this section, the defect correction – a numerical technique to obtain higher-order accurate solutions – is shortly summarized and applied in the FC framework that has been detailed in the previous part.

3.1. Resulting algorithm

A common interest consists in obtaining accurate numerical solutions for comparing them with experimental results. The implicit use of second-order convection–diffusion schemes generally leads to non-M matrices involving difficulties for every iterative linear solver. An alternative may be to employ the defect correction technique (see for example Ref. [18] for mathematical aspects). Noting $\widetilde{\phi}$ a linearizing field, $\mathbf{L}^{\mathbf{1}}_{\widetilde{\phi}}$, $\mathbf{L}^{\mathbf{2}}_{\widetilde{\phi}}$ linear operators, respectively, first-order and second-order accurate and $\mathbf{F}^{\mathbf{1}}_{\widetilde{\phi}}$, $\mathbf{F}^{\mathbf{2}}_{\widetilde{\phi}}$ their associated source terms, the two corresponding linear systems read as follows:

$$\mathbf{L}^{\mathbf{1}}_{\widetilde{\phi}}\,\phi = \mathbf{F}^{\mathbf{1}}_{\widetilde{\phi}}, \qquad \mathbf{L}^{\mathbf{2}}_{\widetilde{\phi}}\,\phi = \mathbf{F}^{\mathbf{2}}_{\widetilde{\phi}}. \tag{19}$$

The defect correction scheme leads to the following iterative form:

$$\underset{\phi}{\mathbf{L}^{\mathbf{1}}}\,\phi = \omega_{\mathrm{DC}}\,(\underset{\phi}{\mathbf{F}^{\mathbf{2}}} - \underset{\phi}{\mathbf{L}^{\mathbf{2}}}\,\widetilde{\phi}) + \underset{\phi}{\mathbf{L}^{\mathbf{1}}}\,\widetilde{\phi}. \tag{20}$$

The idea consists in using an implicit first-order easy-to-invert operator (yielding generally good convergence properties) to obtain a second-order accurate solution. The higher-order discretization scheme is then seen as an outer iteration, which corresponds to a modification of the source term. By the application of the defect correction technique (20), the FC system is sketched in Algorithm 1, where each indexation corresponds to the presumed accuracy of the convection-diffusion scheme (overtilded variables stem from the previous non-linear iteration). In this chart, MFC^{1} denotes the FC method employed with a first-order convection-diffusion scheme. Note that $\mathrm{Nit}_{\mathrm{DC}}$ non-linear iterations of the MFC^{1} FC method are realized to obtain an initial linearizing field for the defect correction method.

Algorithm 1: FC formulation with defect correction technique

(0) Obtain an initial field for the defect correction process:

$$(\mathbf{X}, K, \phi) = \mathrm{MFC}^{1}(\mathbf{X}_0, K_0, \phi_0, \mathrm{Nit}_{\mathrm{DC}}).$$

(1) Update the linearizing field:

$$(\mathbf{X}_0, K_0, \phi_0) = (\mathbf{X}, K, \phi).$$

(2) Resolution of the fully coupled system:

$$\begin{bmatrix} \mathbf{I} & -\mathbf{C}^{\mathbf{1}} & \mathbf{O} \\ -\mathbf{I} & \mathbf{I} & \mathbf{K}_{\mathbf{p}}^{\mathbf{1}}\mathbf{G} \\ \mathbf{D} & \mathbf{O} & -\mathbf{D}\mathbf{K}_{\mathbf{p}}^{\mathbf{1}}\mathbf{G} \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{U}}^{\star} \\ \mathbf{U}^{\star} \\ \mathbf{P}^{\star} \end{bmatrix} = \begin{bmatrix} \omega_{\mathrm{DC}}(\underset{\widehat{\mathbf{U}}}{\mathbf{S}^{\mathbf{2}}} - \widetilde{\widehat{\mathbf{U}}} + \mathbf{C}^{2}\widetilde{\mathbf{U}}) + \widetilde{\widehat{\mathbf{U}}} - \mathbf{C}^{\mathbf{1}}\widetilde{\mathbf{U}} \\ \omega_{\mathrm{DC}}(\widetilde{\widehat{\mathbf{U}}} - \widetilde{\mathbf{U}} - \mathbf{K}_{\mathbf{p}}^{\mathbf{2}}\mathbf{G}\widetilde{\mathbf{P}}) - \widetilde{\widehat{\mathbf{U}}} + \widetilde{\mathbf{U}} + \mathbf{K}_{\mathbf{p}}^{\mathbf{1}}\mathbf{G}\widetilde{\mathbf{P}} \\ \omega_{\mathrm{DC}}(\underset{\mathbf{P}}{\mathbf{S}^{\mathbf{2}}} - \mathbf{D}\widetilde{\widehat{\mathbf{U}}} + \mathbf{D}\mathbf{K}_{\mathbf{p}}^{\mathbf{2}}\mathbf{G}\widetilde{\mathbf{P}}) + \mathbf{D}\widetilde{\widehat{\mathbf{U}}} - \mathbf{D}\mathbf{K}_{\mathbf{p}}^{\mathbf{1}}\mathbf{G}\widetilde{\mathbf{P}} \end{bmatrix}$$

$$\left(\text{until } \frac{\|R_C\|}{\|R_C^{(0)}\|} \leqslant \mathrm{tol}_C\right).$$

(3) Under-relaxation of variables:

$$\mathbf{X} = (1-\omega_C)\mathbf{X}_0 + \omega_C\mathbf{X}^{\star}.$$

(4) Resolution of turbulence equations and under-relaxation:

$$(\mathbf{E}_{\mathbf{K}} - \mathbf{Q}_{\mathbf{K}})K^{\star} = S_K \quad \left(\text{until } \frac{\|R_K\|}{\|R_K^{(0)}\|} \leqslant \mathrm{tol}_K\right),$$

$$(\mathbf{E}_{\phi} - \mathbf{Q}_{\phi})\phi^{\star} = S_{\phi} \quad \left(\text{until } \frac{\|R_{\phi}\|}{\|R_{\phi}^{(0)}\|} \leqslant \mathrm{tol}_{\phi}\right),$$

$$K = (1-\omega_K)K_0 + \omega_K K^{\star},$$

$$\phi = (1-\omega_{\phi})\phi_0 + \omega_{\phi}\phi^{\star}.$$

(5) Repeat steps (1) to (4) until non-linear convergence:

$$\frac{\|R_{\mathrm{NL}}\|}{\|R_{\mathrm{NL}}^{(0)}\|} \leqslant \mathrm{tol}_{\mathrm{NL}}.$$

3.1.1. First- and higher-order discretization schemes

A class of higher-order discretization schemes considered here for operator $\mathbf{C^2}$ is the Van Leer's κ scheme [24]. In this paper, only the QUICK scheme ($\kappa = 1/2$) [25] has been considered. Nevertheless, discretizations with κ-schemes can produce unphysical oscillations near sharp gradients. Thus, total variation diminishing (TVD) schemes have been adopted in order to prevent a solution from oscillating. A recent overview of TVD discretization schemes is given in Ref. [56]. For the applications considered in Part 5, the ISNAS (interpolation scheme which is nonoscillating for advective scalars) limiter [55,56] has been adopted. Concisely, its representation in the $(r, \Psi(r))$ diagram is

$$\Psi(r) = \frac{(|r| + r)(3r + 1)}{2(r + 1)^2}. \tag{21}$$

A standard choice (first-order upwind scheme) is adopted as the first-order convection-diffusion scheme retained in the implicit operator $\mathbf{C^1}$. However, in the defect correction scheme, often the outer iteration controls the convergence speed which can be slow if first and higher-order discretizations are different. This can be justified by reformulating (20) as

$$\phi = \left(\mathbf{I} - \omega_{\mathrm{DC}} \mathbf{L}^{\mathbf{1}^{-1}}_{\underset{\sim}{\phi}} \mathbf{L}^{\mathbf{2}}_{\underset{\sim}{\phi}}\right) \widetilde{\phi} + \omega_{\mathrm{DC}} \mathbf{L}^{\mathbf{1}^{-1}}_{\underset{\sim}{\phi}} \mathbf{F}^{\mathbf{2}}_{\underset{\sim}{\phi}}, \tag{22}$$

where $\mathbf{I}$ is the identity operator, meaning that $\mathbf{L}^{\mathbf{1}}_{\underset{\sim}{\phi}}$ can be seen as a preconditioner of $\mathbf{L}^{\mathbf{2}}_{\underset{\sim}{\phi}}$. Thus, as suggested by Oosterlee et al. [29], it seems more appropriate to retain the "positive" part of $\mathbf{L}^{\mathbf{2}}_{\underset{\sim}{\phi}}$ as implicit operator. This strategy has been tested in Ref. [49] for various TVD schemes. However, the benefits in computational time with respect to the standard choice are not so obvious. A possible explanation is to remark that in the presented applications, the imposed linear residual reduction (tol_C) is equal to 10^{-1} (Algorithm 1, Tables 1 and 4); a more demanding criterion would probably be in favor of the treatment of Oosterlee et al. This aspect needs to be tackled in more detail; but in the following, the standard choice (20) has been adopted.

3.1.2. Construction of the fully coupled system

With a collocated collective ordering of the unknowns such as

$$\mathbf{X}^{\mathbf{T}}_{\mathbf{C}} = (\widehat{\mathbf{U}}, \widehat{\mathbf{V}}, \mathbf{U}, \mathbf{V}, \mathbf{P})_{\mathbf{C}}, \tag{23}$$

it is possible to deduce the following form for the FC system:

$$\mathbf{A_C}(\mathbf{i}, \mathbf{j})\mathbf{X_C} + \sum_{\mathbf{nb}} \mathbf{A_{nb}}(\mathbf{i}, \mathbf{j})\mathbf{X_{nb}} = \mathbf{S_C}(\mathbf{i}, \mathbf{j}), \tag{24}$$

where each block $\mathbf{A_C}(\mathbf{i}, \mathbf{j})$ and $\mathbf{A_{nb}}(\mathbf{i}, \mathbf{j})$ has a 5×5 dimension. This choice leads to a banded block structure for the FC system. The resulting matrix has a symmetric pattern and a block structure.

Table 1
AS240-B airfoil (incidence 19°, $Re = 2 \times 10^6$): parameters for the DC and FC methods

e_1	ω_{vel}	tol_{vel}	ω_P	tol_P	ω_{turb}	tol_{turb}	Scheme (first-order)	Scheme (second-order)
DC method								
50	0.4	10^{-6}	0.3	10^{-3}	0.4	10^{-6}	Upwind	QUICK
	ω_C	tol_C	ω_{turb}	tol_{turb}	Nit_{DC}	ω_{DC}		
FC method								
0	0.8	10^{-1}	0.8	10^{-6}	15	0.8	Upwind	QUICK-ISNAS

The blocks are generally sparse, and their number depends on the retained choices for the discretization schemes. With a classical first-order convection-diffusion scheme, the FC matrix has a global block pentadiagonal structure. The use of higher-order accurate discretization schemes generally leads to a growth of the discretization stencil. Thus, the FC system has now a multi-diagonal structure with at most 13 bands of blocks for two-dimensional applications. Generic forms of blocks are given in Ref. [49] for various schemes.

4. Solution strategies

A new approach has been developed to obtain by an FC procedure higher-order acccurate solutions of the Reynolds-averaged Navier–Stokes equations. The resulting block-banded linear system is neither symmetric nor positive definite. Moreover, when simulating flows on highly stretched curvilinear grids, the poor conditioning and the possible high non-normality may be serious drawbacks. At this stage, the efficiency of the global method will be entirely driven by the efficiency and robustness of the iterative linear solver. In this section, the retained choices will be briefly evoked.

4.1. Linear convergence acceleration techniques

The block-banded linear system requires a reliable solution method to yield a global efficient solution strategy. A key-point is to design a robust and computationally efficient approach. Standard geometric linear multigrid methods are widespread solution strategies of optimal complexity (O(N) operations required for system of N unknowns) [4,18,52]. Optimal efficiency have been proved for problems resulting in M matrices. In the fully coupled framework, the conditions to fulfill the M matrix property for the FC linear system are quite delicate to estimate. Nevertheless, numerical experiments show that on highly stretched curvilinear grids, the smoothing phases handled by classical iterative linear methods (Jacobi, Gauss-Seidel, incomplete lower–upper (ILU) relaxation) are quite inefficient, thus affecting the rate of convergence. Moreover, when targeting the simulation of turbulent flows, their application to the FC system becomes far from trivial. At such Reynolds numbers, as shown by local mode analysis, the convective mechanism is rather difficult to represent on coarse grids, leading to discrepancies in

rate of convergence. A nice cure would be to consider non-standard multigrid methods with multiple semi-coarsening for the linear system of equations to yield grid-independent rates (see Washio et al.'s experience for convection-dominated convection-diffusion or anisotropic diffusion partial differential equations [51]). Algebraic multigrid methods [35] can also be mentioned but are still restricted with respect to parallelism.

Subspace methods or more restrictively Krylov subspace methods are to be preferred because they are applicable to a wide range of problems (i.e., a wide range of matrices). A good overview of Krylov subspace methods is proposed by Barett et al. [3]. Hybrid Bi-CG methods are very attractive due to their low and fixed memory cost, a practical advantage when solving large-scale matrices. Among others, BiCGSTAB [50] is generally adopted. However, the convergence of BiCGSTAB may stagnate. One explanation is the possible inaccuracy on the BiCG coefficients in finite precision arithmetics, that may seriously deteriorate the speed of convergence. As numerical experiments confirm, this happens if the considered (preconditioned) linear system is real and has nonreal eigenvalues with imaginary part that are large relative to the real part [43]. Several cures have been recently proposed [41–43]. In the following, an enhanced hybrid BiCG Krylov subspace method denoted BiCGSTAB-QMR [6] is retained. It combines a quasi-minimization procedure to smooth the Petrov-Galerkin residual and a modification of the minimization (ω parameter) due to Sleijpen et al. [42] to avoid stagnation. In this work, a block-ILU preconditioner is adopted. The construction of the preconditioner $\mathbf{C}$ is as follows:

$$\mathbf{A} = \mathbf{C} - \mathbf{R} \qquad \text{with } \mathbf{C} = (\mathbf{L} + \mathbf{D})\mathbf{D}^{-1}(\mathbf{D} + \mathbf{U}), \tag{25}$$

where $\mathbf{L}$ and $\mathbf{U}$ are respectively lower and upper block parts of $\mathbf{A}$. This preconditioning step induces to construct a diagonal block $\mathbf{D}$ and its full inverse block $\mathbf{D}^{-1}$ with 25 elements per grid point. This choice seems a reasonable compromise between cheap preconditioners like Jacobi, SSOR or polynomial preconditioning which may be inefficient for strongly ill-conditioned matrices and more advanced preconditioners including thresholding, multicoloring-based, multilevel- or domain decomposition-based preconditioners, that need a more pronounced computational effort and memory storage.

An extensive study between various block-ILU preconditioned Krylov subspace methods (including GMRES(m) [36]) has been performed to compare their efficiency in Ref. [49]. As a result, BiCGSTAB-QMR yields the best compromise with respect to computing times, efficiency and memory requirements on a sequential computer.

5. Numerical experiments

A previous publication [11] focused on the simulation of laminar recirculating flows by FC methods in steady or unsteady regime. The classical two-dimensional square lid-driven cavity flow was analyzed at $\mathrm{Re} = 400$ and $\mathrm{Re} = 1000$, whereas three-dimensional calculations concerned the steady cubic lid-driven cavity flow at $\mathrm{Re} = 100$, $\mathrm{Re} = 400$ and $\mathrm{Re} = 1000$ and the unsteady 3:1:1 (spanwise aspect ratio) lid-driven cavity flow at $\mathrm{Re} = 3200$. Spatial accuracy of the FC method with various discretization schemes (including the CPI variant) was analyzed on these configurations. See Ref. [11] for more details. Thus complex two-dimensional fluid flow problems are now treated to continue to investigate the robustness of the FC method. Two steady turbulent flows

will be considered: the separated flow around an AS240-B airfoil at high incidence (19°, $Re = 2 \times 10^6$) and the wake flow behind an hill ($Re = 60\,000$). Applications were carried out on a SGI workstation in double precision arithmetics. A comparison between numerical and experimental results will be proposed for both applications. Besides these comparisons, the goal is twofold:

- assessing and examining the numerical quality of the solution (accuracy and grid-independence of viscous flow predictions)
- evaluating the robustness (or non-robustness) of DC and fully coupled methods on realistic test-cases.

In our opinion, the second point is the most challenging from a numerical point of view. Nevertheless, the question of accuracy and grid-independence of the solution is of relevant interest. About accuracy, it is important to note that DC and FC approaches are only different by the solution phase: in fact the same algebraic pressure–velocity problem is solved by two different routes (a suite of linear systems for the DC method, an unique linear system for the FC one). As a consequence, the whole numerical scheme being identical (discretization schemes, flux reconstruction step, turbulence model, etc.), the results of both methods should be identical at convergence (more precisely at the same level of nonlinear residual reduction). This statement is illustrated in [49] for the simulation of the laminar flow around an hill ($Re = 600$), where the nonlinear residuals can be driven to more than eight orders of reduction. The analysis of physical quantities (separation and reattachment lengths, pressure, velocity profiles) showed that the results of both approaches were found to be identical. The next question that arises is whether it is possible for DC and FC approaches to drive the non-linear residuals to such high level of reduction when simulating turbulent flows. This is the main challenge to be examined in this section.

5.1. Flow past an airfoil at high incidence

5.1.1. Goal and configuration

The accurate prediction of forces and moments on an airfoil at high incidence represents a quite challenging task from a numerical point of view due to strong physical difficulties. These limitations (turbulence modeling, transition, scale effects) are noteworthy in stall and post-stall situations where the flow induces massive separation. A detailed numerical study has been already proposed by Guilmineau et al. [17] on various airfoils in pre-stall and post-stall conditions, where both numerical and physical detailed comments and bibliographic references are given. Here only static stall calculations on a commercial airfoil shape (AS 240-B) will be performed. The flow past this airfoil at 19° of incidence ($Re: 2 \times 10^6$) has been experimentally studied at ONERA-CERT in Toulouse by Gleyzes [16]. It has been proved that the flow is fully turbulent, stalled and can be considered as two-dimensional in the mean. The AS240-B airfoil has a 16% thickness involving a very progressive stall and a velocity overshoot close to the leading edge, low enough to prevent immediate transition.

5.1.2. Numerical aspects

Mesh generation is handled by a conformal transformation method, leading to O-type curvilinear orthogonal grids with high stretching (mesh A: 128×128, mesh B: 192×128). First points

away from the wall correspond to a dimensionless value of y^+ about 0.5. The outer flow boundary is located at 13 chords. On the outer boundary, a uniform velocity field $U = U_\infty$ is imposed. On the airfoil, no slip conditions are prescribed.

Three formulations will be compared:

1. DC method: The classical PISO DC formulation [21] with ILU preconditioned Krylov subspace method (BiCGSTAB) as a linear solver for each variable.
2. DC-MG: The same DC formulation with ILU preconditioned BiCGSTAB as the linear solver for velocity and turbulence variables and a linear multigrid solver for the elliptic pressure problem. The cell-centered linear multigrid method uses as a smoother ILU(β) method [54] and a Galerkin coarse grid approximation with arithmetic restriction and linear prolongation. F-cycles with one pre- and post-smoothing steps are adopted within a five level hierarchy for mesh A.
3. FC method: This method (with defect correction technique) with block ILU-preconditioned BiCGSTAB-QMR as linear solver for the FC system, ILU preconditioned BiCGSTAB for the turbulence variables K and ω.

The various parameters for the three algorithms are gathered in Table 1. The parameters in the upper part of this table represent standard choices for DC methods: ω_{vel} represents the under-relaxation parameters for the velocity variables ($\omega_{\text{vel}} = \omega_U = \omega_V$), in the same way $\omega_{\text{turb}} = \omega_K = \omega_\omega$. Note that a strong percentage of diagonal dominance (controlled by the e_1 parameter) is mandatory (see Ref. [12] for more details). The pressure linear reduction criterion (tol_P) has been chosen after extensive numerical experiments to test its reliability.

In the lower part of this table, note that no diagonal dominance is activated ($e_1 = 0$) in the FC procedure, whereas quite large under-relaxation parameters are imposed. This may lead to a fast nonlinear convergence if and only if the solution phase is successful. The fact that no approximate inverse is involved in the fully implicit procedure is here exploited. This notably contrasts to the standard DC approaches. Finally, the parameters Nit_{DC} and ω_{DC} involved in the defect correction approach are given.

Non-linear convergence analysis: Table 2 collects the required number of nonlinear iterations (Nit) to reach convergence, the corresponding spent CPU time per point (σ) in s/point on meshes A and B for the different methods. Fig. 2 (upper part) presents the corresponding nonlinear convergence histories. In ordinate is represented the evolution of the normalized residual in logarithmic law, that is, if r denotes a generic residual, $r^{(0)}$ an initial residual, the plot of the

Table 2
AS240-B airfoil (incidence 19°, $\text{Re} = 2 \times 10^6$): analysis of the non-linear convergence for DC-based methods, DC and DC-MG and FC method

Mesh	Method					
	DC		DC-MG		FC	
	Nit	σ (s/point)	Nit	σ (s/point)	Nit	σ (s/point)
A: 128×128	44000	7.18	44000	5.5	1555	0.47
B: 192×128	62000	5.96	–	–	1133	0.571

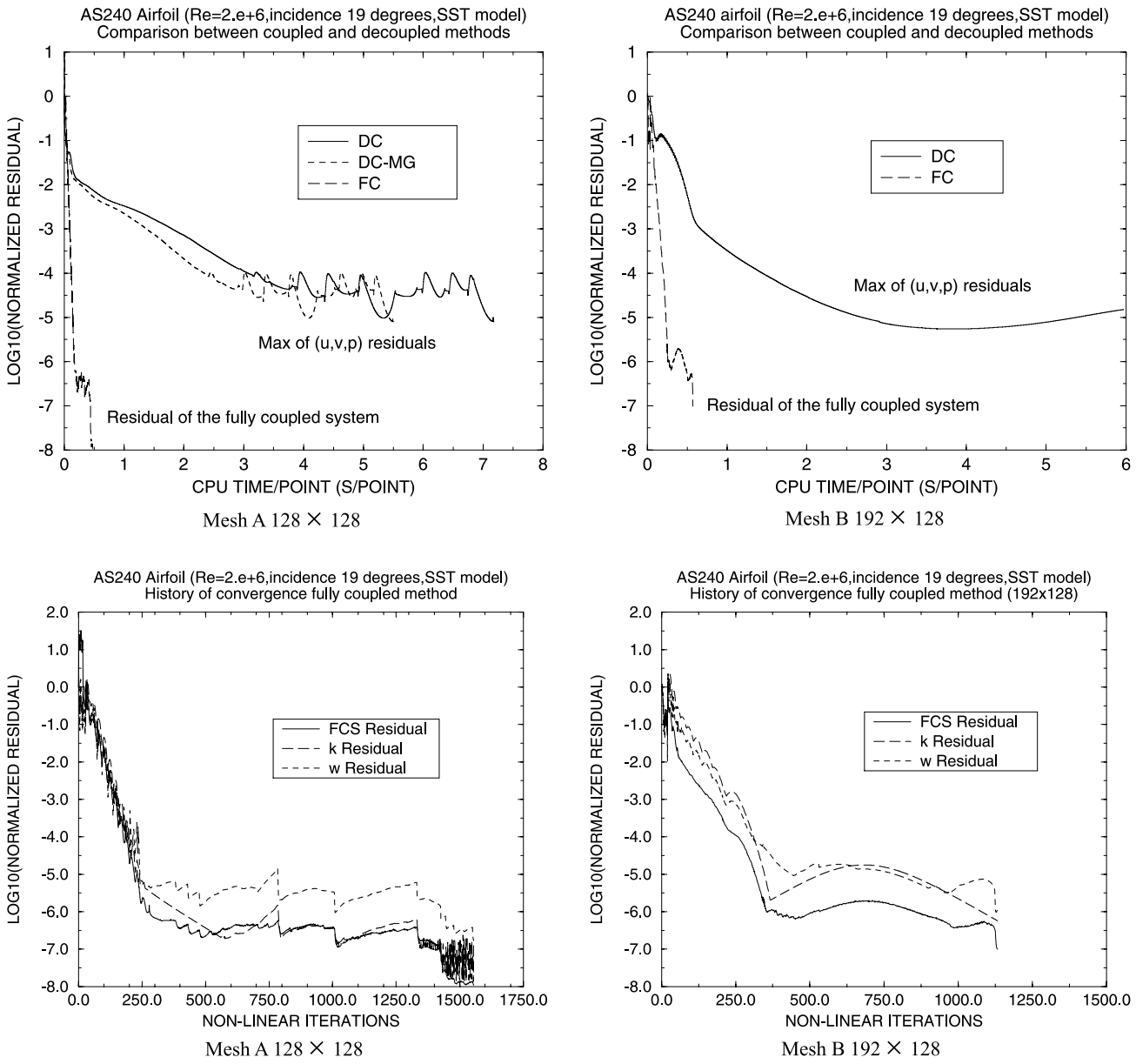


Fig. 2. AS240-B profile (incidence: 19°, $\mathrm{Re} = 2 \times 10^6$): history of the nonlinear convergence for DC and FC methods on meshes A and B with QUICK as second-order accurate discretization scheme and K–ω SST as turbulence model.

function $\log_{10}(\|r\|_2/\|r^{(0)}\|_2)$ where $\|\ \|_2$ designes the l_2-norm. This choice easily allows a comparison of different methods for the same level of normalized residual reduction. For the ease of presentation, the residual of the fully coupled system is compared with the maximum of the pressure–velocity residuals stemming from the DC methods. Lower part of Fig. 2 shows the complete non-linear convergence history of the FC method on both meshes (note the different abscisse with respect to the upper part).

Fig. 2 shows a classical behavior of nonlinear convergence for DC, when simulating turbulent flows on moderately complex geometries. The convergence rate is quite fast in the beginning of the

nonlinear process, then a pronounced slope breakdown appears, ended by either oscillations or stagnation phenomena. The nonlinear reduction seems to reach a limit cycle (corresponding to five orders of reduction approximately), whereas oscillations (on mesh A) are difficult to interpret. A fixed number of nonlinear iterations has been therefore set up on both meshes (Table 1). As expected, the behavior of DC-MG is similar. The multigrid method cuts down the CPU time spent in the pressure solver by a factor greater than two. Nevertheless this satisfactory result induces a weak global computational gain with respect to DC (a factor close to 1.3). A similar (or slighty better) result would have been obtained on mesh B.

Above all, the four curves of Fig. 2 reveal the dramatic improvement due to the implicit pressure–velocity coupling. Moreover, this coupled procedure does not exhibit some stalling behavior in the non-linear process and leads to an efficient method with respect to CPU time. For example a residual reduction of four orders on the coarse mesh A requires a computational effort equal to 3.3 s/point, 2.51 s/point and 0.115 s/point, respectively, for the DC, DC-MG and FC approaches. Thus the FC approach leads to a computational gain equal to 28.7 and 21.8 with respect to DC and DC-MG for this case. The same comparison on the finer mesh B leads to a computational gain about 10 with respect to DC. These results are extremely attractive regarding the complexity of the flow. By using a new algorithmic approach, a computational time speed-up corresponding to a *four or five years* evolution in hardware technology has been reached. Note also that, on this test-case, besides efficiency, the FC method can be considered as robust, since the non-linear reduction can be driven to seven or eight orders of magnitude without any stalling phenomena.

Linear convergence analysis: For both simulations, the number of preconditioned BiCGSTAB-QMR linear iterations to fulfill the linear residual reduction criterion (see Algorithm 1):

$$\frac{\|R_C\|}{\|R_C^{(0)}\|} \leqslant \mathrm{tol}_C \quad \text{with} \quad \mathrm{tol}_C = 10^{-1} \tag{26}$$

is represented during the non-linear process in Fig. 3. This evolution illustrates the robustness of the FC method. A common property is found: at the beginning of the non-linear process, a rather high number of iterations is mandatory. Up to 98 and 206 linear iterations at most are mandatory respectively on meshes A and B. Note also the severe variations in linear iterations. This step corresponds to the main nonlinear reduction phase (from 0 to 6 orders of magnitude). Then, a phase follows where a small number of iterations is needed. This behavior can be explained by analysis of Fig. 2, where each plateau in Fig. 3 corresponds to a stagnation in non-linear convergence. To avoid this stagnation, a more demanding linear residual reduction criterion would be mandatory, allowing a possible decrease in nonlinear iterations. An adaptive strategy for estimating this criterion would be probably helpful.

5.1.3. Physical aspects

For the considered turbulence model and second-order discretization schemes, the FC method allows to obtain a solution converged at the "engineer's accuracy" in a satisfying complexity on this example. The main interest is now to analyse the quality of the numerical solution.

Pressure coefficient: Table 3 presents the results obtained on meshes A and B for DC and FC approaches. Note that the experimental drag coefficient is not available. Results obtained with

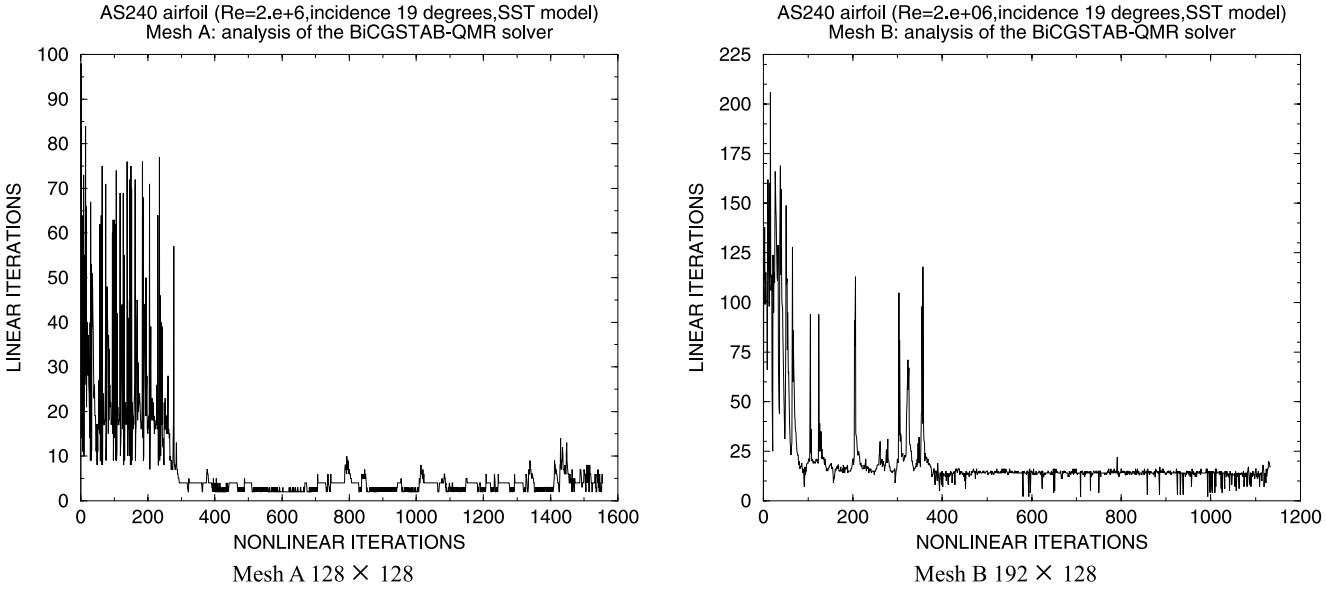


Fig. 3. AS240-B profile (incidence: 19°, $\mathrm{Re} = 2 \times 10^6$): analysis of the behavior of the linear solver (BiCGSTAB-QMR) on meshes A and B with QUICK as second-order accurate discretization scheme and K–ω SST as turbulence model.

Table 3
AS240-B airfoil (incidence 19°, $\mathrm{Re} = 2 \times 10^6$): integral coefficients (C_D: drag coefficient, C_L lift coefficient)[a]

	Mesh	Mesh					
	Mesh	A: 128 × 128			B: 192 × 128		
	Scheme	QUICK		ISNAS	QUICK		ISNAS
	EXP	DC	FC	FC	DC	FC	FC
C_D	Not available	0.166	0.13684	0.141	0.134	0.13123	0.132
C_L	1.268	1.3309	1.307	1.273	1.311	1.293	1.30

[a] Comparison between experimental and numerical results from the DC and FC methods ($K-\omega$ SST model).

two second-order discretization schemes have been used for the FC method (QUICK and ISNAS), while only results with the QUICK scheme are presented for the DC one. The corresponding evolution of the pressure coefficient is given in Fig. 4 for the FC method. QUICK and ISNAS only differ by the level of the suction peak on both meshes. As expected, ISNAS leads to a smoother peak which is nevertheless still overpredicted. With QUICK as second-order accurate discretization scheme, Fig. 5 proposes a direct comparison on the pressure coefficient between DC and FC results on the upper part, while the lower part shows for each approach (DC or FC) the discrepancies between solutions obtained on meshes A and B. Note that DC and DC-MG lead to the same results, since DC-MG just involves an improvement on the convergence of the pressure linear system. Fig. 5 puts in light the same trends for DC and FC methods. The discrepancies

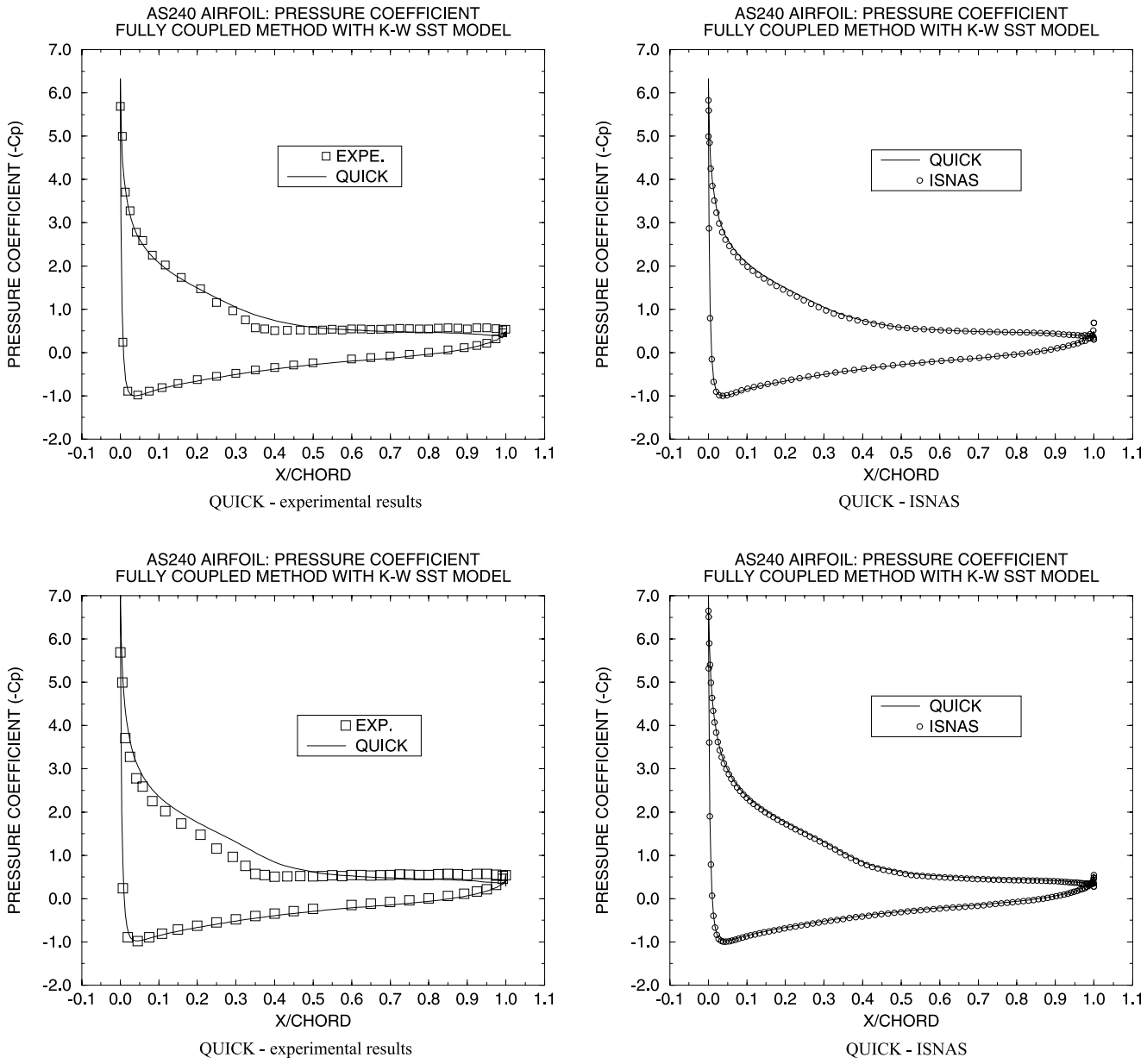


Fig. 4. AS240-B profile (incidence: 19°, Re $= 2 \times 10^6$, mesh A 128×128 (up), mesh B 192×128 (down)): comparison of the pressure coefficient between experimental results [16] and numerical results issued from the FC method (on the left), influence of the second-order accurate discretization scheme on the pressure coefficient (on the right).

shared by both approaches mainly concern the level of suction peak and the separation phase. In the pressure plateau zone, FC solutions match well contrary to DC ones. These differences are mainly a consequence to the different level of reduction of the nonlinear residuals. As a general comment, the accuracy of the pressure plateau and the suction peak appears rather moderate. The suction peak is at 6.2 on the coarse mesh, while close to 7 on the finer one. After numerical experiments, TVD schemes yield almost identical results on the plateau accuracy and differ only by

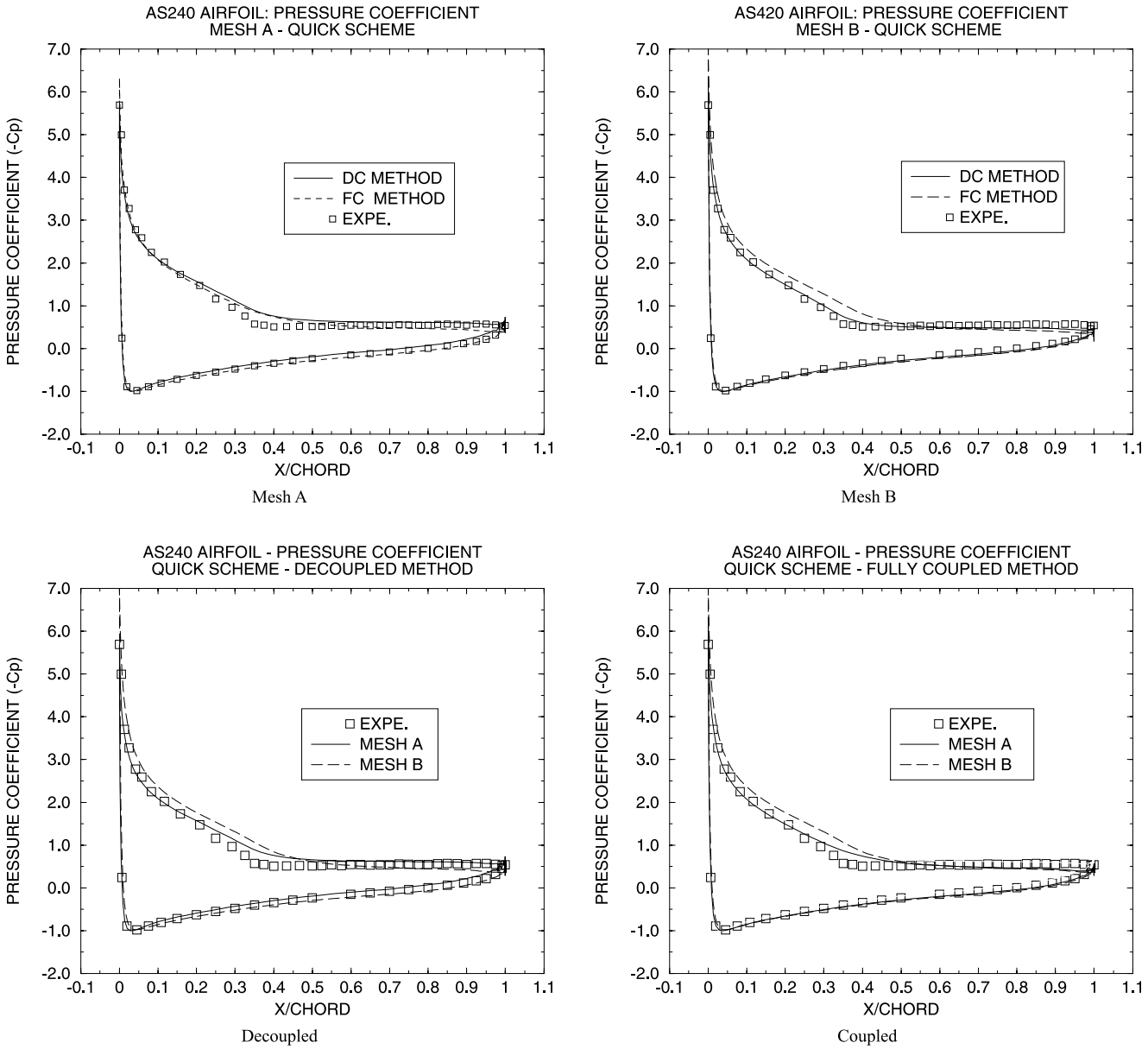


Fig. 5. AS240-B profile (incidence: 19°, $Re = 2 \times 10^6$): comparison of the pressure coefficient between experimental results [16] and numerical results issued from the FC and DC methods.

the level of the suction peak (only very slight differences have been observed). According to our numerical experiments, these discrepancies are mainly felt to be a consequence of the adopted higher-order accurate convection-diffusion discretization scheme. As a possible proof, this configuration has also been studied in Ref. [17] within a DC framework including the CPI closure and various turbulence models (including the K–ω SST variant. Quite similar trends for the level of suction peak (overprediction) were found. Note that mesh-independent solutions were found in [17] on conformal meshes of size 200×90 and 400×180. Table 3 and Figs. 4 and 5 point out that a further study on finer meshes should be undertaken to conclude.

5.2. Wake flow behind a hill

5.2.1. Goal and configuration

The wake flow behind a two-dimensional hill presents strong favorable and adverse pressure gradients, streamline curvature with stabilizing and destabilizing effects on turbulence and very high velocity gradients in the near-wall region on the top of the hill. The problem is of relevant interest since it involves numerous physical phenomena: separation, recirculation and reattachment. Moreover this configuration experimentally studied by Almeida [1,2] has been set up for the Fourth ERCOFTAC-IAHR Workshop on Refined Flow Modelling organized by Bonnin, Buchal and Rodi [34].

The channel height is $H = 170$ mm, the height and length of the hill are $h = 28$ mm, $l = 108$ mm. This flow has been measured by laser techniques [1,2] at $\mathrm{Re} = 60\,000$, where the Reynolds number is based on the centerline mean velocity $U_0 = 2.47$ m/s and the hill height. Velocity and kinetic energy profiles are determined at several sections down the hill.

The inlet boundary is located 100 mm upstream of the center of the hill. The inlet boundary conditions for the velocity and the turbulent kinetic energy are prescribed according to experimental data. The specific dissipation rate $\omega = \sqrt{k}/l$ is deduced from the length scale l:

$$l = \min(\kappa y, 0.05H), \quad 0 < y < \frac{H}{2}, \tag{27}$$

$$l = \min(\kappa(H - y), 0.05H), \quad \frac{H}{2} < y < H. \tag{28}$$

5.2.2. Numerical aspects

Mesh generation is handled by a transfinite interpolation yielding nonorthogonal and highly stretched grids (mesh A: 118×74, mesh B: 234×146). Only two formulations will be considered for this configuration: the DC and the FC methods. Used parameters are gathered in Table 4 (same notations as in Table 1).

Non-linear convergence analysis: The results of both formulations (DC and FC) are gathered in Table 5. Once again, a maximal number of non-linear iterations has been set up for the decoupled formulation. This is clearly explained by analysis of the convergence curve, Fig. 6, where

Table 4
Wake flow behind a hill ($\mathrm{Re} = 60\,000$): parameters for the DC and FC methods

e_1	ω_{vel}	$\mathrm{tol}_{\mathrm{vel}}$	ω_{P}	$\mathrm{tol}_{\mathrm{P}}$	ω_{turb}	$\mathrm{tol}_{\mathrm{turb}}$	Scheme (first-order)	Scheme (second-order)
Hill – DC method								
50	0.35	10^{-6}	0.35	10^{-3}	0.4	10^{-6}	Upwind	QUICK
	ω_C	tol_C	ω_{turb}	$\mathrm{tol}_{\mathrm{turb}}$	$\mathrm{Nit}_{\mathrm{DC}}$	ω_{DC}		
Hill – FC method								
0	0.8	10^{-1}	0.8	10^{-6}	15	0.8	Upwind	QUICK-ISNAS

Table 5
Wake flow behind a hill ($Re = 60\,000$, K–ω model of Wilcox): analysis of the non-linear convergence on different meshes for DC and FC method [a]

Mesh	Method							
	DC				FC			
	Nit	σ (ms/point)	x_s/h	x_r/h	Nit	σ (ms/point)	x_s/h	x_r/h
A: 118×74	2500	330.65	0.22	6.58	154	64.2	0.21	6.45
B: 234×146	4000	688.44	0.22	6.72	135	143.5	0.214	6.61

[a] Separation and reattachment lengths (QUICK scheme).

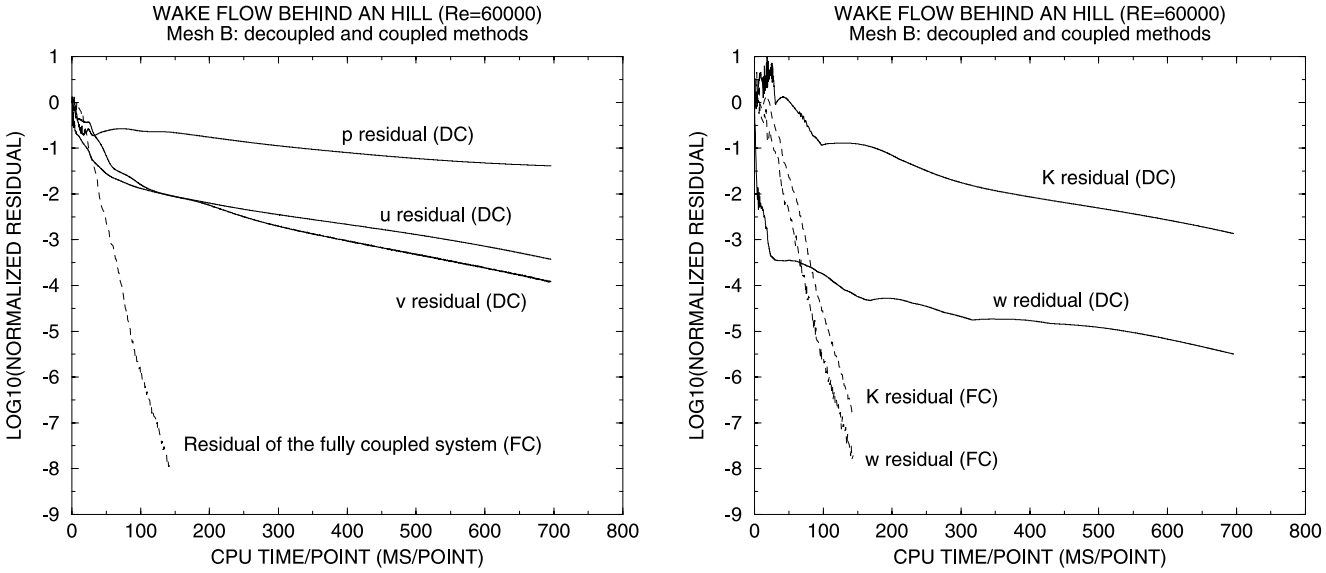


Fig. 6. Hill ($Re = 60\,000$, mesh B 234×146): history of the non-linear convergence for DC and FC methods with QUICK as second-order accurate discretization scheme and K–ω (Wilcox) as turbulence model.

the non-linear convergence of each variable is detailed for the DC formulation. Fig. 6 also provides a comparison of histories of convergence between the two formulations (DC and FC). The nonlinear process is not slowed down by a limit-cycle behavior of the FC method, confirming its interest. Note that this remark has been established on a very fine grid (234×146).

Linear convergence analysis: As in the previous part, the linear convergence behavior of the block ILU-preconditioned BiCGSTAB-QMR is presented (Fig. 7). For this test-case, the linear solver does not show any phase of "easy solving" that induces oscillations in the non-linear process. This is confirmed in Fig. 6.

5.2.3. Physical aspects

Separation and reattachment parameters: The physical analysis concerns both the separation and reattachment parameters. Table 5 sums up the results for both formulations. A good

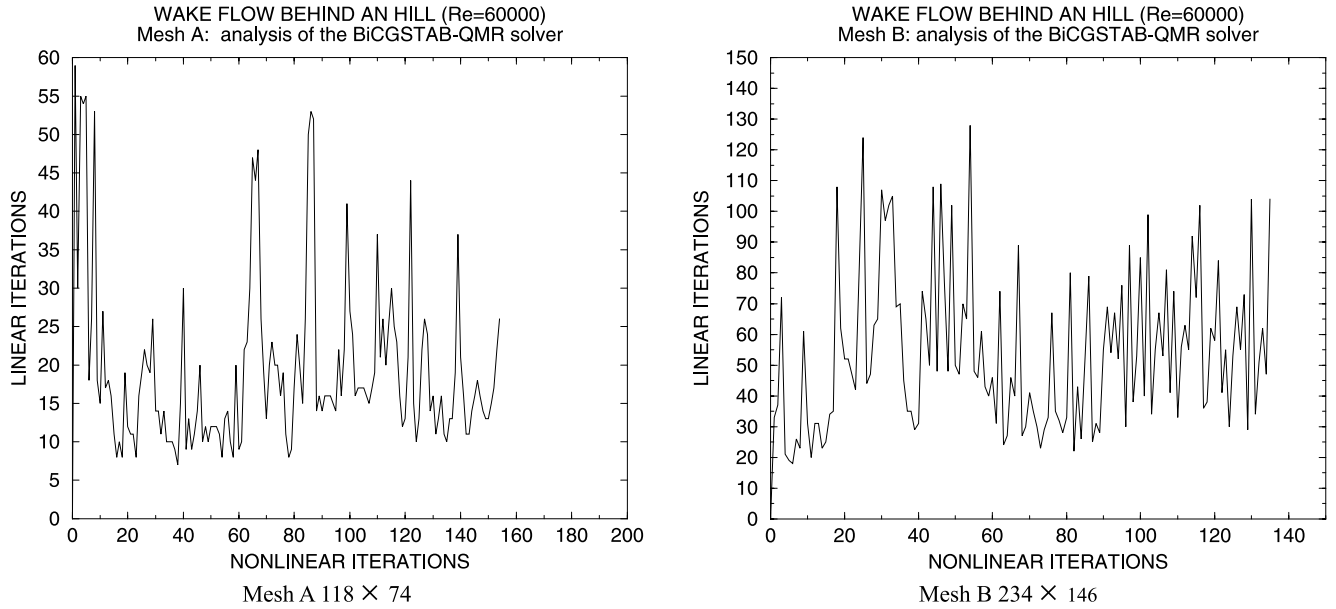


Fig. 7. Hill (Re = 60 000): analysis of the behavior of the linear solver (BiCGSTAB-QMR) on meshes A and B with QUICK as second-order accurate discretization scheme and K–ω (Wilcox) as turbulence model.

agreement is found on the separation length. Nevertheless the prediction of the reattachment length is mesh-dependent for both methods. The comparison between DC and FC results is rather delicate. As outlined before, Fig. 6 shows that both solutions have not reached the same level of non-linear residual reduction. As a consequence, discrepancies are expected. According to our experiments, the solution method being fixed, the differences on recirculation length between meshes A and B are attributed to the turbulence model [37]. This statement has also been drawn during the ERCOFTAC workshop (see Ref. [56] for a discussion), where a comparison of various results of numerous teams shows an overprediction of 43% for the separation length. Present results are comparable.

Profiles: A comparison of velocity, kinetic energy profiles is proposed at different sections located downstream the hill (Figs. 8–10). The presented results are obtained with the FC method. About FC method, the comparison of the profiles on meshes A and B shows a good agreement between the solutions, notably for the prediction of the u-component and the kinetic energy. Besides the profiles (Fig. 9), where some differences are noticed, the solution can be considered as mesh-independent, at least for the considered sections. The differences between DC and FC solutions are weak: only Figs. 9 and 10 (middle and bottom) show some slight differences in the peak zones. The comparison with experimental results seems to indicate a deficiency of the K–ω of Wilcox in the prediction of normal velocity components. The analysis of the profiles (Fig. 10) shows that the K–ω (Wilcox model) captures the main effects but generally leads to an overestimation of the longitudinal velocity profiles upstream of separation and to a correct capture of the separated zone. However vertical velocity profiles are strongly overestimated. The turbulent kinetic energy peaks, although well located, are strongly underestimated by about 40%. Similar comments and conclusions were drawn in the ERCOFTAC workshop [56]. The overall prediction displays a too slow recovery to fully developed channel flow.

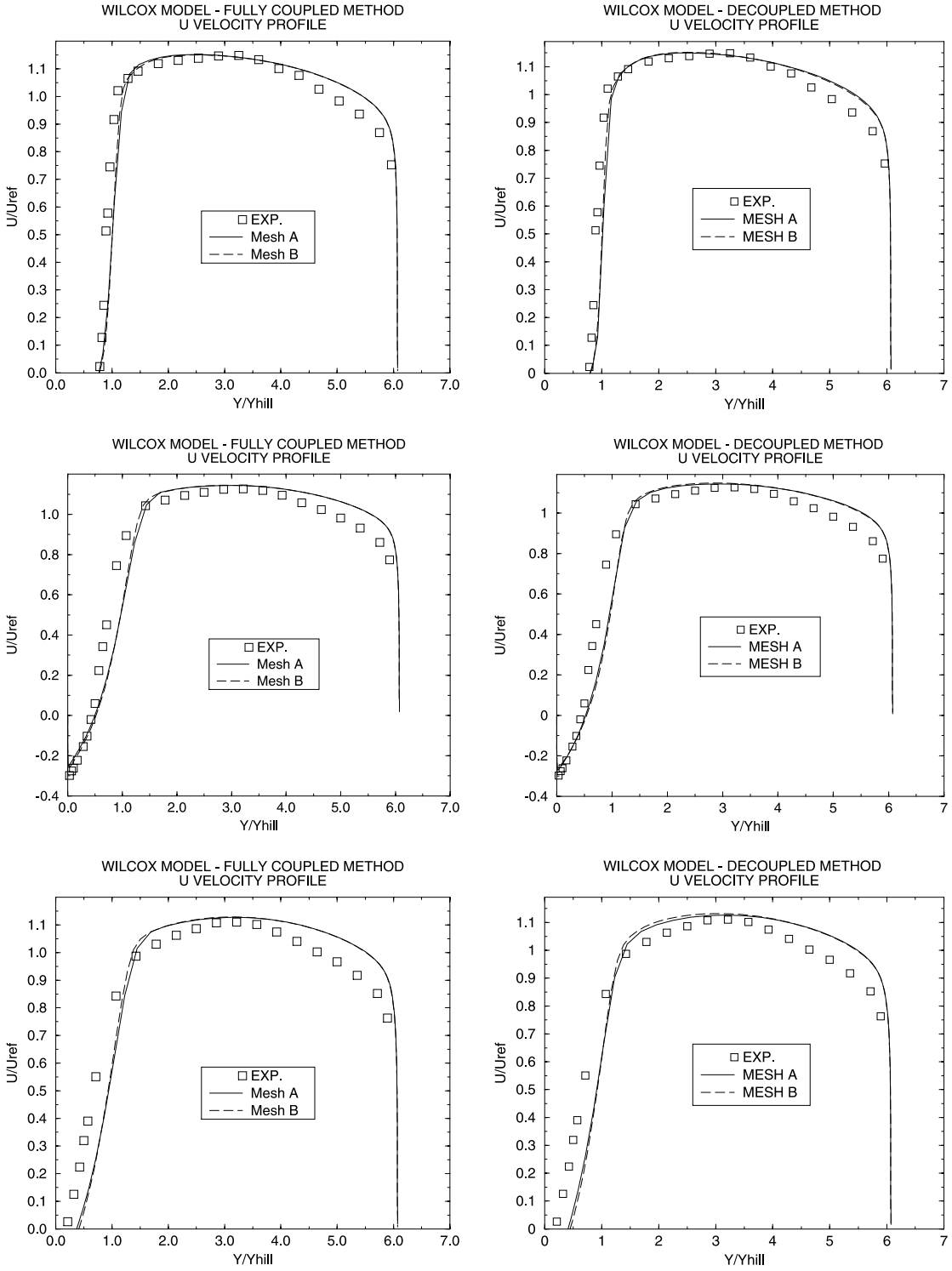


Fig. 8. Comparison on streamwise velocity profiles at $x = 0.03$ m (up), $x = 0.09$ (middle) and $x = 0.12$ m (bottom) for FC (left) and DC (right) methods.

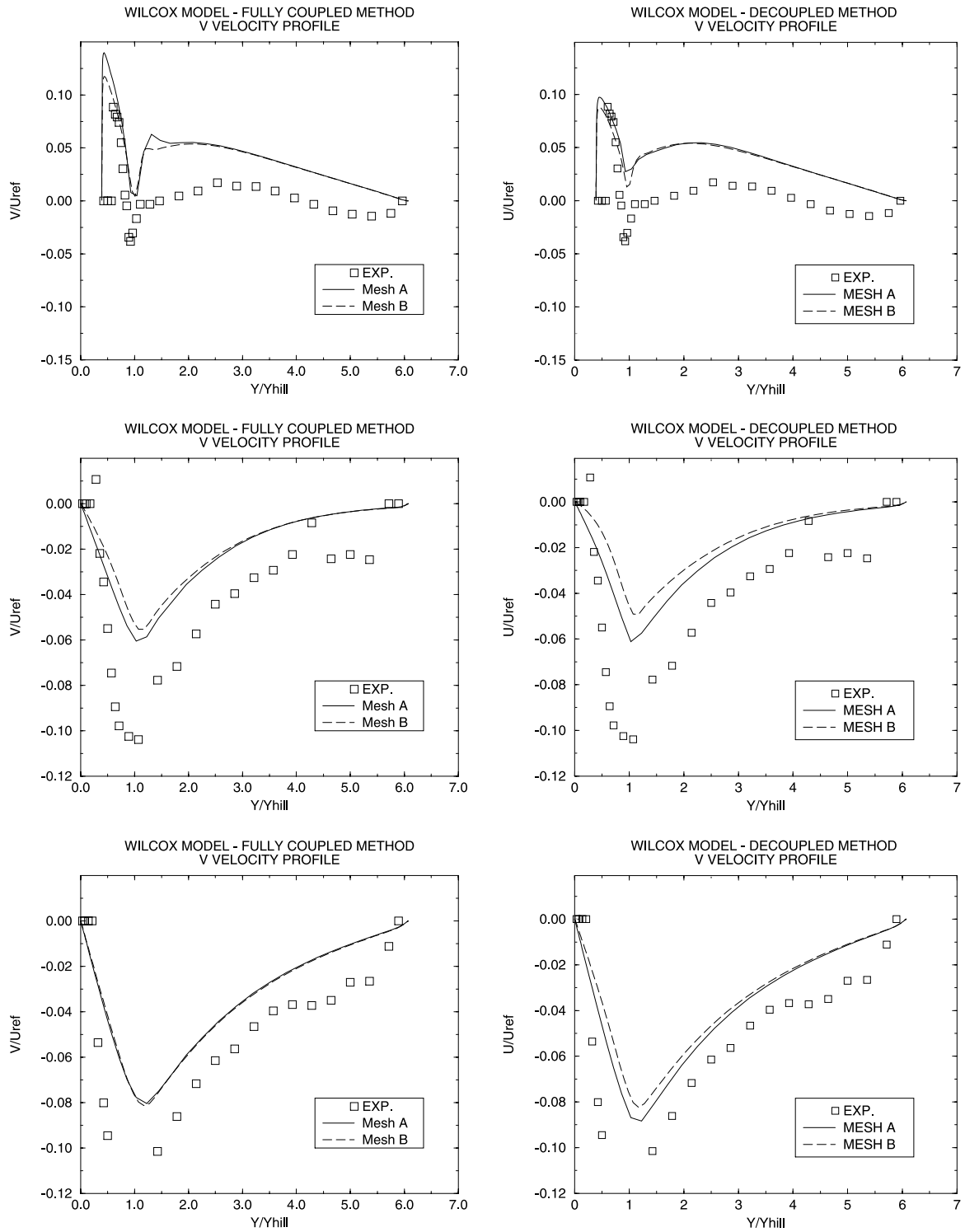


Fig. 9. Comparison on vertical velocity profiles at $x = 0.03$ m (up), $x = 0.09$ (middle) and $x = 0.12$ m (bottom) for FC (left) and DC (right) methods.

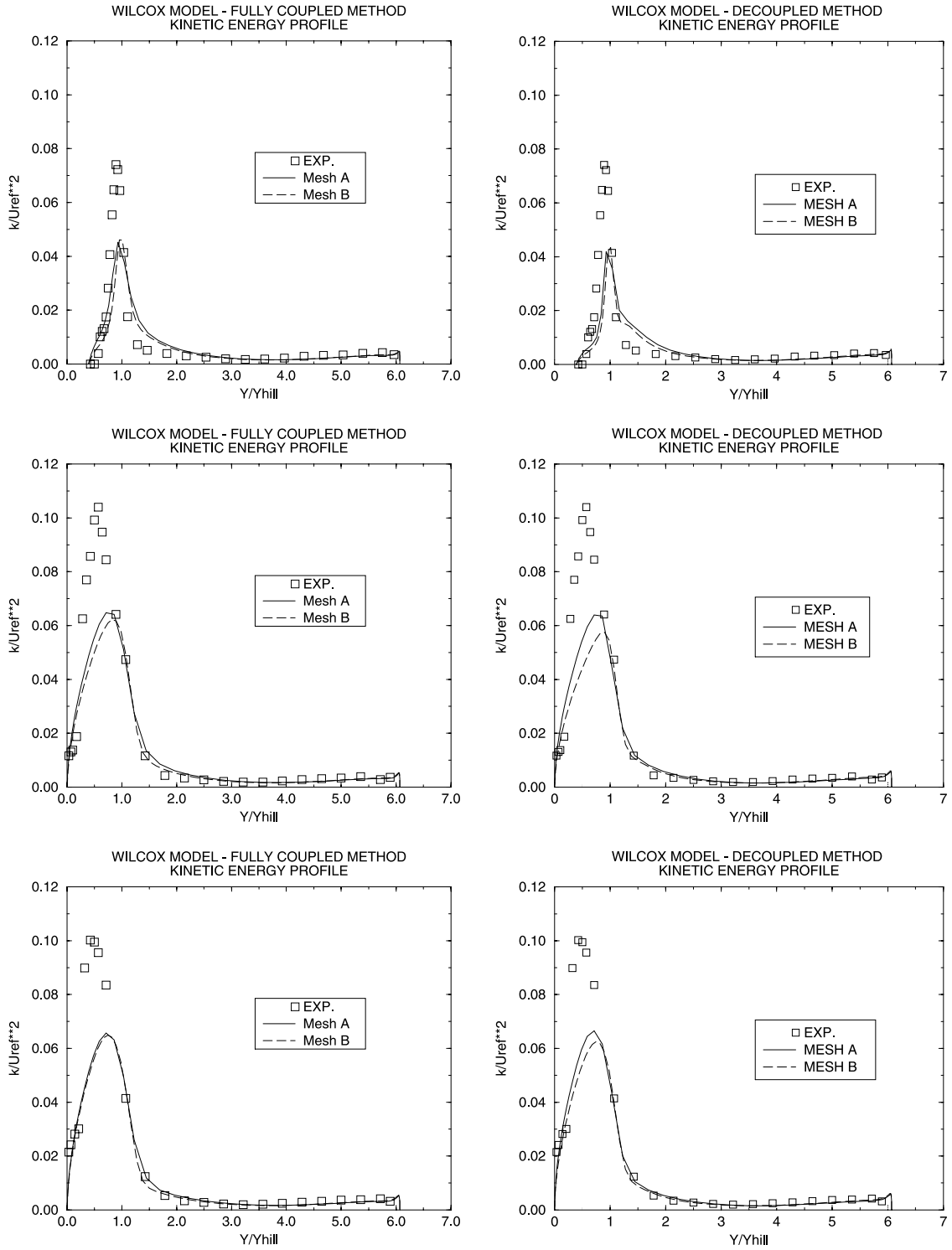


Fig. 10. Comparison on turbulent kinetic energy profiles at $x = 0.03$ m (up), $x = 0.09$ (middle) and $x = 0.12$ m (bottom) for FC (left) and DC (right) methods.

6. Concluding remarks and perspectives

A FC procedure has been proposed for solving the incompressible Reynolds-averaged Navier–Stokes equations on moderately complex geometries. This strategy initially proposed in Ref. [9] has been "revisited" in order both to obtain more accurate numerical solutions and to enhance the solution phase, the true keystone part. The former point has been treated with the defect correction technique, a practical and popular way that consists in a source term modification. Its implementation and the choice of first- and higher-order discretization schemes have been detailed. The coupled system of algebraic equations induces a strongly ill-conditioned, non symmetric, very large linear system, on refined stretched curvilinear grids. Therefore the major limitation of the FC procedure consists in finding and designing a computationally efficient and robust linear solver. A block ILU preconditioned hybrid BiCG method has been used yielding satisfying convergence behavior, at least for the considered applications.

To estimate the potential of the new approach, the numerical simulation of turbulent flows involving physical phenomena such as separation, reattachment or recirculation has been investigated by different methodologies (either DC- or FC-based). The analysis of the numerical results proves the robustness of the proposed FC method. The global and implicit treatment of the pressure–velocity coupling is the main reason for this increased robustness. As a proof, the non-linear residuals of the FC system do not exhibit any limit cycling behavior in contrast to DC methods. Both treated examples show a non-linear reduction of seven or eight orders of magnitude for a truly modest CPU time on a single workstation. The numerical quality of the computational solutions has been examined on various grid sizes for DC and FC methods. This analysis induces two main comments. First, the comparison between DC and FC solutions is delicate on these test-cases and can be misleading. Due to the different non-linear convergence behavior, both solutions correspond to different level of non-linear residual reduction. Discrepancies are therefore expected. Second, the comparison between numerical and experimental solutions yields two distinct comments. The simulation of the flow past an airfoil at high incidence has shown that numerical results (obtained either by DC or FC methods) are mesh-dependent. This is believed to be due to the chosen discretization schemes. A cure has been proposed in Ref. [17] with a new flux closure. As already discussed, it would be interesting to evaluate such fully coupled formulation with the CPI closure on such demanding test-case. The simulation of the turbulent flow behind a two-dimensional hill has shown that the results can be considered as mesh-independent. Nevertheless, discrepancies with experimental results can be important (especially for the turbulent kinetic energy). As already pointed out in Ref. [56], this is believed to be due to the two-equation turbulence modelling.

Some perspectives of development can be briefly evoked. To sum up, they are twofold. First, the need of a new flux reconstruction step has been outlined to circumvent the introduction of the pseudo-velocity variables. The goal is to decrease the size of the FC system, a nice feature when targetting large-scale applications [11]. Secondly, the adoption of non-linear acceleration techniques (either multigrid- or subspace-based) seems attractive to still improve the robustness of the FC method. See Ref. [49] for some first applications. Besides these possible enhancements, the simulation of turbulent flows around three-dimensional moderately complex geometries by such a FC procedure seems even more challenging. This project is currently under development and should lead to a robust and computationally efficient solution method. To study large-scale

problems (note that a three-dimensional mesh of $n_1 n_2 n_3$ control volumes induces a FC system of size $7n_1n_2n_3 \times 7n_1n_2n_3$), a parallel version of the code should be designed through grid partitioning and message passing. At long range, such parallel Reynolds-averaged Navier–Stokes solver could be inserted in an optimization loop, due to its expected numerical robustness and computational efficiency.

Remark: For the sake of brevity, some elements (algorithms, meshes, configurations, views of the flow) that the reader may find interesting are available on request to the authors.

References

[1] Almeida GP, Durao DFG, Heitor MV. Wake flows behind two-dimensional model hills. Exp Thermal Fluid Sci 1993;7:87.
[2] Almeida GP, Durao DFG, Heitor MV, Simoes JP. LDV measurements of fully-developed turbulent channel flow. Proc Fifth Int Symp Appl Laser Tech Fluid Mech, Lisbon, 1990. p. 9–12.
[3] Barrett R, Berry M, Chan T, Demmel J, Donato J, Dongarra J, Eijkhout V, Pozo R, Romine C, van der Vorst H. Templates for the solution of linear systems: Building Blocks for Iterative Methods. SIAM, Philadelphia, 1994.
[4] Brandt A. Multi-level adaptive solutions for boundary value problems. Math Comput 1977;31:333–90.
[5] Brown P, Saad Y. Convergence theory of nonlinear Newton–Krylov algorithms. SIAM J Optim 1994;4:297–330.
[6] Chan TF, Gallopoulos E, Simoncini V, Szeto T, Tong CH. A quasi-minimal residual variant of the BICGSTAB algorithm for nonsymmetric systems. SIAM J Sci Stat Comput 1994;15:338–47.
[7] Clift SS, Forsyth PA. Linear and non-linear iterative methods for the incompressible Navier–Stokes equations. Int J Numer Meth Fluids 1994;18:229–56.
[8] Kevin Cope Wm, Vanka SP, Wang G. Multigrid calculations of twin jet impingement with crossflow: comparison of segregated and coupled relaxation strategies. ASME Fluids Engineering Summer Annual Meeting, Lake Tahoe, NV, June 19–23, 1994.
[9] Deng GB, Ferry M, Piquet J, Visonneau M. New fully coupled solutions of the Navier–Stokes equations. Notes Numer Fluid Mech 1991;35:191–200.
[10] Deng GB, Piquet J, Queutey P, Visonneau M. Incompressible flow calculations with a consistent physical interpolation using the CPI method. Comput Fluids 1994;23(8):1020–47.
[11] Deng GB, Piquet J, Queutey P, Visonneau M. A new fully coupled solution of the Navier–Stokes equations. Int J Numer Meth Fluids 1994;19:605–39.
[12] Deng GB, Piquet J, Queutey P, Visonneau M. Navier–Stokes equations for incompressible flows: finite-difference and finite volume methods. In Peyret R, editor. Handbook of Computational Fluid Mech, New York: Academic Press; 1996, p. 25–97.
[13] Deng GB, Piquet J, Visonneau M. Viscous flow computations using a fully coupled technique. Proc. Second International Colloquium on viscous fluid dynamics in ship and ocean technology, Osaka, 1991, p. 186–202.
[14] Devine KD, Hennigan GL, Hutchinson SA, Salinger AG, Shadid JN, Tuminaro RS. High performance MP unstructured finite element simulation of chemically reacting flows. Conference on SuperComputing , SC'97, San Jose, Nov 15–21 1997.
[15] Dick E, Linden J. A multigrid method for steady incompressible Navier–Stokes equations based on flux difference splitting. Int J Numer Meth Fluids 1992;14:1311–23.
[16] Gleyzes C. Opération décrochage – résultats d'essais á la soufflerie F2. Technical Report OA 71/2259 AYD (DERAT 55/5004-22), ONERA-DERAT, 1988.
[17] Guilmineau E, Piquet J, Queutey P. Two-dimensional turbulent viscous flow simulation past airfoils at fixed incidence. Comput Fluids 1997;26:135–62.
[18] Hackbusch W. Multigrid methods and applications. Springer-Verlag 1985.
[19] Hanby RF, Silvester DJ, Chew JW. A comparison of coupled and segregated iterative solution techniques for incompressible swirling flow. Int J Numer Meth Fluids 1996;22:353–73.

[20] Hutchinson SA, Shadid J, Tuminaro R. Aztec user's guide. Technical Report SAND95-1559, Sandia National Laboratory, 1995.
[21] Issa R. Solution of the implicitly discretized fluid flows equations by operator-splitting. J Comput Phys 1986;62(1):40–65.
[22] Karki KC, Mongia HC. Evaluation of a coupled solution approach for fluid flow calculations in body-fitted coordinates. Int J Numer Meth Fluids 1990;11:1–20.
[23] Knoll D, MacHugh P. Enhanced non-linear iterative techniques applied to a non-equilibrium plasma flow. SIAM J Sci Comput 1998;19(1):291–301.
[24] Van B. Upwind-difference methods for aerodynamic problems governed by the Euler equations. Lect Appl Math 1985;22:327–36.
[25] Leonard BP. A stable and accurate convective modelling procedure based on quadratic upstream interpolation. Comput Meth Appl Mech Engng 1979;19:59–98.
[26] MacHugh P, Knoll D. Fully coupled finite volume solutions of the incompressible Navier–Stokes and energy equations using an inexact Newton method. Int J Numer Meth Fluids 1994;19:439–55.
[27] Menter FR. Zonal two-equations k–ω turbulence models for aerodynamic flows. AIAA 24th Fluid Dynamics Conference, AIAA Paper-93-2906, Orlando, 1993.
[28] Oosterlee CW. Robust multigrid methods for the steady and unsteady Navier–Stokes equations in general coordinates. PhD Thesis, University of Technology, Department of Mathematics and Informatics, Delft, 1993.
[29] Oosterlee CW, Gaspar F, Washio T, Wienands R. Multigrid line smoothers for higher order upwind discretizations of convection-dominated problems. J Comput Phys 1998;139(2):274–307.
[30] Patankar S. Numerical heat transfer and fluid flow. Hemisphere Publishing Corporation, 1980.
[31] Patankar S, Spalding D. A calculation procedure for heat, mass and momentum transfer in three-dimensional parabolic flows. Int J Heat Mass Transfer 1972;15:1787–806.
[32] Pernice M, Walker H. NITSOL: a Newton iterative solver for nonlinear systems. SIAM J Sci Comput 1998;19(1):302–18.
[33] Rhie C, Chow W. A numerical study of the turbulent flow past an isolated airfoil with trailing edge separation. AIAA J 1983;21:1525–32.
[34] Rodi W, Bonnin JC, Buchal T. ERCOFTAC workshop on data bases and testing of calculation methods for turbulent flows. In Proc. Fourth ERCOFTAC-IAHR Workshop on Refined Flow Modelling, Karlsruhe, Germany, 1995.
[35] Ruge J, Stüben K. Algebraic multigrid. In S.Mc Cormick, editor. Multigrid Methods, SIAM Philadelphia, 1987, p. 73–130.
[36] Saad Y, Schultz M. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J Sci Stat Comput 1986;7:856–69.
[37] Schmid M, Deng GB, Seidl V, Visonneau M, Peric M. Computation of complex turbulent flows. In: E.H. Hirschel, editor. Numerical flow simulation I, Notes on numerical fluid mechanics, Vol. 66. Braunschweig: Vieweg-Verlag, 1998, p. 407–25.
[38] Schneider GE, Zedan M. A coupled modified strongly implicit procedure for the numerical solution of coupled continuum problems. Proc. AIAA/SAE/ASME 20th Joint Propulsion Conference, AIAA-84-1743, Cincinnati, 1984.
[39] Shadid J, Moffat H, Hutchinson SA, Hennigan G, Devine K, Salinger A. MPSalsa: a finite element computer program for reacting flow problems part I: Theorical development. Technical Report SAND96-2752, Sandia National Laboratory, 1996.
[40] Shadid JN, Tuminaro RS, Walker HF. An inexact Newton method for fully-coupled solution of the Navier–Stokes equations with heat and mass transport. J Comput Phys 1997;137(1):155–85.
[41] Sleijpen GLG, Fokkema DR. BICGSTAB (l) for linear equations involving unsymmetric matrices with complex spectrum. Electron Trans Numer Anal 1993;1:11–32.
[42] Sleijpen GLG, van der Vorst HA. Maintaining convergence properties of BICGSTAB methods in finite precision arithmetic. Numer Algorithms 1995;10:203–23.
[43] Sleijpen GLG, van der Vorst HA. An overview of approaches for the stable computation of hybrid Bi-CG methods. Appl Numer Math 1996;19:235–54.

[44] Sockol P. Multigrid solution of the Navier–Stokes equations on highly stretched grids. Int J Numer Meth Fluids 1993;17:543–66.
[45] Thompson MC, Ferziger JH. An adaptive multigrid technique for the incompressible Navier–Stokes equations. J Comput Phys 1989;82:94–121.
[46] Vanka SP. Block implicit coupled calculation of internal fluid flow. In Fifth Symposium on Turbulent Shear Flows, Cornell University, 1985, p. 20–d32.
[47] Vanka SP. Block-implicit multigrid solution of Navier–Stokes equations in primitive variables. J Comput Phys 1986;65:138–58.
[48] Varga RS. Matrix iterative analysis. Englewood Cliffs, NJ: Prentice-Hall; 1962.
[49] Vasseur X. Etude numérique de techniques d'accélération de convergence lors de la résolution des équations de Navier–Stokes en formulation découplée ou fortement couplée. PhD Thesis, Université de Nantes, 1998.
[50] van der Vorst HA. BICGSTAB: a fast and smoothly converging variant of BiCG for the solution of nonsymmetric linear systems. SIAM J Sci Stat Comput 1992;13(2):631–44.
[51] Washio T, Oosterlee CW. Flexible multiple semicoarsening for three-dimensional singularly perturbed problems. SIAM J Sci Comput 1998;19(5):1646–66.
[52] Wesseling P. An introduction to multigrid methods. Chichester: Wiley; 1992.
[53] Wilcox DC. Turbulence modeling for CFD DCW Industries, California, USA, 1993.
[54] Wittum G. On the robustness of ILU-smoothing. In W.Hackbusch, editor, Robust Multigrid Methods, Fourth GAMM-Seminar, Notes on Numer Fluid Mech vol.23, Vieweg, 1988, p. 217–39 .
[55] Zijlema M, Wesseling P. Higher-order flux-limiting schemes for the finite volume computation of incompressible flow. Int J Comput Fluid Dyn 1998;9:89–109.
[56] Marcel Zijlema.Computational modeling of turbulent flow in general domains. PhD Thesis, University of Technology, Department of Mathematics and Informatics, Delft, 1996.

B.2. Domain decomposition preconditioners of Neumann-Neumann type for hp approximations on boundary layer meshes in three dimensions

IMA Journal of Numerical Analysis (2004) **24**, 123–156

Domain decomposition preconditioners of Neumann–Neumann type for *hp*-approximations on boundary layer meshes in three dimensions

ANDREA TOSELLI† AND XAVIER VASSEUR‡
Seminar for Applied Mathematics, ETH Zürich, Rämistrasse 101, CH-8092 Zürich, Switzerland

[Received on 8 January 2003; revised on 15 April 2003]

We develop and analyse Neumann–Neumann methods for *hp* finite-element approximations of scalar elliptic problems on geometrically refined boundary layer meshes in three dimensions. These are meshes that are highly anisotropic where the aspect ratio typically grows exponentially with the polynomial degree. The condition number of our preconditioners is shown to be independent of the aspect ratio of the mesh and of potentially large jumps of the coefficients. In addition, it only grows polylogarithmically with the polynomial degree, as in the case of *p* approximations on shape-regular meshes. This work generalizes our previous one on two-dimensional problems in Toselli & Vasseur (2003a, submitted to *Numerische Mathematik*, 2003c to appear in *Comput. Methods Appl. Mech. Engng.*) and the estimates derived here can be employed to prove condition number bounds for certain types of FETI methods.

Keywords: domain decomposition; preconditioning; *hp* finite elements; spectral elements; anisotropic meshes.

1. Introduction

Solutions of elliptic boundary value problems in polyhedral domains have corner and edge singularities and, in addition, boundary layers may also arise in laminar, viscous, incompressible flows with moderate Reynolds numbers at faces, edges and corners. Suitably graded meshes, geometrically refined towards corners, edges and/or faces, can be employed in order to achieve an exponential rate of convergence of *hp* finite-element approximations (see e.g. Andersson *et al.*, 1995; Babuška & Guo, 1996; Melenk & Schwab, 1998; Schwab & Suri, 1996; Schwab *et al.*, 1998).

Neumann–Neumann (NN) and FETI algorithms are particular iterative substructuring methods and are among the most popular and heavily tested domain decomposition (DD) methods (see e.g. Le Tallec, 1994; Farhat & Roux, 1994; Mandel & Brezina, 1996; Bhardwaj *et al.*, 2000). Unfortunately, the performance of iterative substructuring methods might be severely compromised if very thin elements and/or subdomains or general non-quasiuniform meshes are employed.

Some work has been done on domain decomposition preconditioners for higher-order approximations of three-dimensional problems. It is well-known that on shape-regular

†Email: toselli@sam.math.ethz.ch
‡Email: vasseur@sam.math.ethz.ch

IMA Journal of Numerical Analysis 24(1), © Institute of Mathematics and its Applications 2004; all rights reserved.

meshes special care must be taken in the choice of the basis functions in order to produce preconditioners that are robust with respect to the polynomial degree (e.g. Mandel, 1989, 1990a,b; Pavarino, 1994; Bica, 1997; Sherwin & Casarin, 2001). For p approximations that employ nodal basis functions on Gauss–Lobatto nodes (spectral element approximations), many iterative substructuring methods can be successfully employed and studied (see Pavarino & Widlund, 1996, 1997; Pavarino, 1997; Pavarino & Warburton, 2000 and the references therein). Some of these ideas can be and have been generalized to hp approximations (e.g. Ainsworth, 1996a,b; Oden *et al.*, 1997; Guo & Cao, 1997; Le Tallec & Patra, 1997; Ainsworth & Sherwin, 1999; Korneev *et al.*, 2002 and the references therein and, in particular, Guo & Cao, 1998 for three-dimensional problems). In all the above-mentioned works, however, the finite-element mesh is assumed to be shape-regular and robustness with respect to the aspect ratio is not in general ensured and often unlikely to hold in practice.

In Toselli & Vasseur (2003a,c), we showed that NN and FETI methods can be successfully devised for the particular geometrically refined boundary layer meshes commonly used for hp finite-element approximations of two-dimensional problems. Indeed, these meshes are highly anisotropic, but of a particular type:

1. they are obtained by refining an initial *shape-regular* mesh (macromesh);
2. refinement is only carried *towards* the boundary of the computational domain.

These properties, also shared by three-dimensional meshes, allowed us to obtain condition number bounds for the corresponding preconditioned operators that only grow polylogarithmically with the polynomial degree, as is the case of p approximations on shape-regular meshes. Our understanding and analysis was confirmed by numerical experiments. In particular, we choose the macromesh as a decomposition into substructures in such a way that subdomains are shape-regular. Roughly speaking, the reason why such favourable condition numbers are retained lies in the fact that upper bounds come from stable decompositions of finite-element functions into components associated with geometrical objects (typically vertices and edges of the subdomains in two dimensions). Because of our particular meshes, only components associated with *internal* vertices need to be considered, i.e. relative to vertices in a neighbourhood of which the mesh is shape-regular.

Three-dimensional boundary layer meshes also share the two characteristics mentioned above. However, stable decompositions now involve face and wirebasket components, where the wirebasket is the union of the subdomain edges and vertices that do not lie on the external boundary of the computational domain. By considering, for instance, an edge E of a macroelement that shares a face with Ω (see the face patch in Fig. 1, left, or Fig. 2), decoupling of face and wirebasket components is now also performed close to $\partial\Omega$, and thus where the mesh is not shape-regular. In this work, we are however able to provide condition number bounds that only grow polylogarithmically with the polynomial degree, as in the two-dimensional case, and are independent of arbitrarily large aspect ratios of the mesh.

The core of this work lies in the careful modification and derivation of certain Sobolev-type inequalities that are independent of the aspect ratio of the mesh for wirebasket and face components of finite-element functions; see Section 7. Provided such inequalities are available, the definition of the algorithms and their analysis are fairly standard procedures

in DD methods and proceed as in the two-dimensional case in Toselli & Vasseur (2003a). Here, we will only consider the *balancing method*, which belongs to the family of Neumann–Neumann methods, but note that the estimates derived can be employed for the analysis of other Neumann–Neumann methods and one-level FETI methods in a straightforward way (see Pavarino, 1997; Klawonn & Widlund, 2001; Toselli & Vasseur, 2003a).

We limit our analysis to the case of nodal basis functions built on Gauss–Lobatto nodes. In addition, we only consider the model problem (2.1), which does not have boundary layers but only corner and edge singularities. However, our tensor-product meshes can also be employed when only singularities are present and do not require the use of hanging nodes. We recall that numerical results in Toselli & Vasseur (2003c) for two-dimensional problems showed that better performance is obtained for certain singularly perturbed problems which exhibit boundary layers. In addition, a linear dependence in k for the condition number was observed for problems with geometric refinement towards interfaces that lie in the interior of the computational domain.

The remainder of this paper is organized as follows: in Sections 2 and 3, we introduce our continuous and discrete problems, respectively. Geometric boundary layer meshes are introduced in Section 4. A particular choice of basis functions is given in Section 5 and our Neumann–Neumann preconditioners are defined in Section 6. Section 7 is the core of this work and is devoted to the proof of some discrete Sobolev-type inequalities. Comparison results for certain discrete harmonic extensions are given in Section 8. Condition number bounds are then proven in Section 9. Section 10 contains some numerical results, while some concluding remarks and perspectives are presented in Section 11.

2. Problem setting

We consider a linear, elliptic problem on a bounded polyhedral domain $\Omega \subset \mathbb{R}^3$ of unit diameter, formulated variationally as:
Find $u \in H_0^1(\Omega)$, such that

$$a(u, v) = \int_\Omega \rho(\mathbf{x}) \nabla u \cdot \nabla v \, \mathrm{d}\mathbf{x} = f(v), \quad v \in H_0^1(\Omega). \tag{2.1}$$

As usual, $H^1(\Omega)$ is the space of square summable functions with square summable first derivatives, and $H_0^1(\Omega)$ its subspace of functions that vanish on $\partial\Omega$. The functional $f(\cdot)$ belongs to the dual space $H^{-1}(\Omega)$. Here $\mathbf{x} = (x, y, z)$ denotes the position vector.

The coefficient $\rho(\mathbf{x}) > 0$ can be discontinuous, with very different values for different subregions, but we allow it to vary only moderately within each subregion. We will in fact assume that the region is the union of elements (also called subdomains, substructures, or macroelements) $\{\Omega_i\}$. Without decreasing the generality of our results, we will only consider the piecewise constant case:

$$\rho(\mathbf{x}) = \rho_i, \quad \mathbf{x} \in \Omega_i.$$

In the case of a region of diameter H_i, such as the substructure Ω_i, we use a norm with

different relative weights obtained by a simple dilation argument:

$$\|u\|^2_{1,\Omega_i} = |u|^2_{1,\Omega_i} + \frac{1}{H_i^2}\|u\|^2_{0,\Omega_i}. \tag{2.2}$$

Here, $\|\cdot\|_{0,\Omega_i}$ and $|\cdot|_{1,\Omega_i}$ denote the norm in $L^2(\Omega_i)$ and the seminorm in $H^1(\Omega_i)$, respectively. In the following we also employ the space $W^{1,\infty}(\Omega_i)$ of bounded functions with bounded derivatives (see e.g. Nečas, 1967).

3. *hp* finite-element approximations

We now specify a particular choice of finite-element spaces. Given an affine quadrilateral mesh $\mathcal{T}$ of Ω and a polynomial degree $k \geqslant 1$, we consider the following finite-element spaces:

$$X = X^k(\Omega; \mathcal{T}) := \{u \in H_0^1(\Omega) \mid \quad u_{|K} \in \mathbb{Q}_k(K),\ K \in \mathcal{T}\}. \tag{3.1}$$

Here $\mathbb{Q}_k(K)$ is the space of polynomials of maximum degree k in each variable on K. In the following, we may drop the reference to k, Ω, and/or $\mathcal{T}$ whenever there is no confusion.

In this paper, we always assume that the meshes are *regular*, i.e. the intersection between neighbouring elements is either a vertex, or an edge, or a face that is common to the *two* elements.

A finite-element approximation of (2.1) consists of finding $u \in X$, such that

$$a(u, v) = f(v), \quad v \in X. \tag{3.2}$$

4. Geometric boundary layer meshes

In order to resolve boundary layers and/or singularities, geometrically graded meshes can be employed. They are determined by a mesh grading factor $\sigma \in (0, 1)$ and a refinement level $n \geqslant 0$. The number of layers is $n + 1$ and the thinnest layer has a width proportional to σ^n. Robust exponential convergence of hp finite-element approximations is achieved if n is suitably chosen. For singularity resolution, n is required to be proportional to the polynomial degree k (see Andersson *et al.*, 1995; Babuška & Guo, 1996). For boundary layers, the width of the thinnest layer needs to be comparable to that of the boundary layer (see Melenk & Schwab, 1998; Schwab & Suri, 1996; Schwab *et al.*, 1998).

A geometric boundary layer mesh $\mathcal{T} = \mathcal{T}_{bl}^{n,\sigma}$ is, roughly speaking, the tensor product of meshes that are geometrically refined towards the faces. Figure 1 shows the construction of a geometric boundary layer mesh $\mathcal{T}_{bl}^{n,\sigma}$.

The mesh $\mathcal{T}_{bl}^{n,\sigma}$ is built by first considering an initial shape-regular macro-triangulation $\mathcal{T}_m$, possibly consisting of just one element, which is successively refined. This process is illustrated in Fig. 1. Every macroelement can be refined isotropically (not shown) or anisotropically in order to obtain so-called face, edge or corner patches (Fig. 1, level 2). Here and in the following, we only consider patches obtained by triangulating the reference cube $\hat{Q} := I^3$, with $I := (-1, 1)$. A patch for an element $K_m \in \mathcal{T}_m$ is obtained by using an affine mapping $F_{K_m} : \hat{Q} \to K_m$. The stability properties proven for patches on the

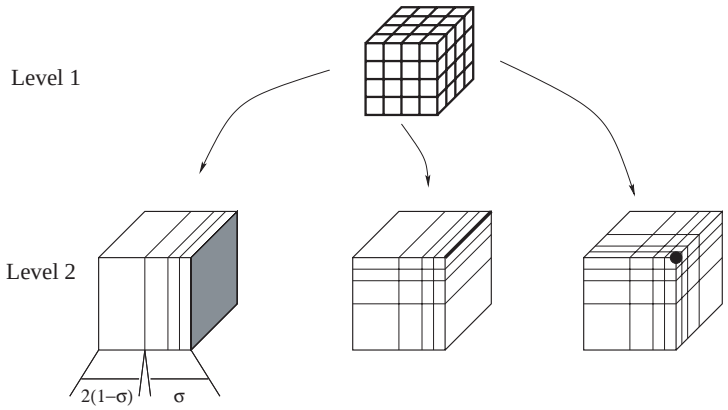


FIG. 1. Hierarchic structure of a boundary layer mesh, with $\sigma = 0{\cdot}5$ and $n = 3$.

reference cube are equally valid for an arbitrary shape-regular element $K_m \in \mathcal{T}_m$, with a constant that is independent of the diameter of K_m.

A *face patch* is given by an anisotropic triangulation of the form

$$\mathcal{T}_f := \{K_x \times I \times I \mid \quad K_x \in \mathcal{T}_x\}, \tag{4.1}$$

where $\mathcal{T}_x$ is a mesh of I, geometrically refined towards, say, $x = 1$, with grading factor $\sigma \in (0, 1)$ and n levels of refinement; see Fig. 1 (level 2, left). We note that the mesh $\mathcal{T}_x \times \{I\}$ of $\hat{S} := I^2$ is a two-dimensional edge patch.

An *edge patch* is given by a triangulation

$$\mathcal{T}_e = \mathcal{T}_e^{bl} := \{K_x \times K_y \times I \mid \quad K_x \in \mathcal{T}_x,\ K_y \in \mathcal{T}_y\} = \{K_{xy} \times I \mid \quad K_{xy} \in \mathcal{T}_{xy}\}, \tag{4.2}$$

where $\mathcal{T}_x$ and $\mathcal{T}_y$ are meshes of I, geometrically refined towards, say, $x = 1$ and $y = 1$, respectively, with grading factor $\sigma \in (0, 1)$ and total number of layers n; see Fig. 1 (level 2, centre). The mesh $\mathcal{T}_{xy}$ of $\hat{S}$ is a two-dimensional corner patch.

In a similar way, we can define a *corner patch* $\mathcal{T}_c$:

$$\mathcal{T}_c = \mathcal{T}_c^{bl} := \{K_x \times K_y \times K_z \mid \quad K_x \in \mathcal{T}_x,\ K_y \in \mathcal{T}_y,\ K_z \in \mathcal{T}_z\},$$

where $\mathcal{T}_x$, $\mathcal{T}_y$, and $\mathcal{T}_z$ are meshes of I, geometrically refined towards, say, $x = 1$, $y = 1$, and $z = 1$, respectively; see Fig. 1 (level 2, right).

We note that every element $\hat{K}$ of $\mathcal{T}_f$, $\mathcal{T}_e$, and $\mathcal{T}_c$ on the reference cube is of the form $(0, h_x) \times (0, h_y) \times (0, h_z)$ (after a possible translation and rotation) and is thus obtained from the reference element by an affine mapping $F_{\hat{K}} : \ \hat{Q} \to \hat{K}$ of the form

$$[x \quad y \quad z]^T = [(h_x/2)(\hat{x} + 1) \quad (h_y/2)(\hat{y} + 1) \quad (h_z/2)(\hat{z} + 1)]^T. \tag{4.3}$$

The aspect ratio of $\hat{K}$ is the maximum of all possible ratios of h_x, h_y and h_z. Since the macromesh consists of affinely mapped elements K_m, every element K of the global mesh $\mathcal{T} = \mathcal{T}_{bl}^{n,\sigma}$ is obtained from the reference element by combining two affine mappings

$$K = F_K(\hat{Q}) = F_{K_m}(F_{\hat{K}}(\hat{Q})), \quad K \subset K_m \in \mathcal{T}_m. \tag{4.4}$$

Since $\mathcal{T}_m$ is shape-regular, the aspect ratio is determined only by $F_{\hat{K}}$; cf. (4.3). Finally, we note that the aspect ratio of the mesh is determined by σ and n, and is proportional to σ^{-n}.

As in Toselli & Vasseur (2003a), our analysis will be made for a prototype mesh, obtained from a shape-regular (not necessarily quasi-uniform) macromesh, by refining elements that only touch $\partial\Omega$, either as corner, edge, or face patches. Such meshes only consist of four types of patches: unrefined, face, edge, and corner patches. We also recall that in practical applications σ is bounded away from one and zero.

5. Basis functions on Gauss–Lobatto nodes

For the space $X^k(\Omega; \mathcal{T})$, we choose nodal basis functions on the Gauss–Lobatto nodes. We denote by $GLL(k)$ the set of Gauss–Lobatto points $\{\xi_i;\ 0 \leqslant i \leqslant k\}$ on $I = (-1, 1)$ in increasing order and by $\{w_i > 0\}$ the corresponding weights (see Bernardi & Maday, 1997, Section 4). We recall that the quadrature formula based on $GLL(k)$ has order $2k-1$ and, in addition,

$$\|u\|_{0,I}^2 \leqslant \sum_{i=0}^{k} u(\xi_i)^2\, w_i \leqslant 3\,\|u\|_{0,I}^2, \quad u \in \mathbb{Q}_k(I); \tag{5.1}$$

(see Bernardi & Maday, 1997, Remark 13.3).

For the reference cube $\hat{Q} = (-1, 1)^3$ we set $GLL(k)^3 = \{\xi_{ijl} = (\xi_i, \xi_j, \xi_l);\ 0 \leqslant i, j, l \leqslant k\}$. In the following, we use the same notation for the mapped Gauss–Lobatto nodes and corresponding weights for an affinely mapped element $K \in \mathcal{T}$.

Given the nodes $GLL(k)^3$, our basis functions on $\mathbb{Q}_k(\hat{Q})$ are the tensor product of kth-order Lagrange interpolating polynomials on $GLL(k)$, defined by

$$\hat{l}_i(\xi_j) = \delta_{ij}. \tag{5.2}$$

On the reference element we can write

$$u(x, y, z) = \sum_{i=0}^{k}\sum_{j=0}^{k}\sum_{l=0}^{k} u(\xi_i, \xi_j, \xi_l)\ \hat{l}_i(x)\hat{l}_j(y)\hat{l}_l(z), \quad u \in \mathbb{Q}_k(\hat{Q}). \tag{5.3}$$

For a general element in $\mathcal{T}$, basis functions are obtained by mapping those on the reference element. *Interior* local basis functions correspond to GLL nodes inside $\hat{Q}$ (all local indices differ from 0 and k).

Equation (5.3) defines an interpolation operator I^k on the reference element

$$I^k u(x, y, z) := \sum_{i=0}^{k}\sum_{j=0}^{k}\sum_{l=0}^{k} u(\xi_i, \xi_j, \xi_l)\ \hat{l}_i(x)\hat{l}_j(y)\hat{l}_l(z).$$

The points $GLL(k)^3$ define a triangulation $\mathcal{T}_k = \mathcal{T}_k(\hat{Q})$ of $\hat{Q}$ in a natural way, consisting of k^3 parallelepipeds. Let $Y^h = Y^h(\hat{Q}) = X^1(\hat{Q}; \mathcal{T}_k)$ be the space of piecewise trilinear functions on this mesh. We also denote $Y^k = Y^k(\hat{Q}) = \mathbb{Q}_k(\hat{Q})$. The aspect ratio of $\mathcal{T}_k$ is of the order of k (see Casarin, 1996, p. 27 for details). In a similar way we can consider a Gauss–Lobatto mesh on an affinely mapped element K by simply mapping the

GLL mesh on $\hat{Q}$. In the following, we will use the notation $\mathcal{T}_k = \mathcal{T}_k(K)$, $Y^h = Y^h(K)$ and $Y^k = Y^k(K)$ to denote the GLL mesh, the piecewise trilinear finite-element space and $\mathbb{Q}_k$, respectively, for a mapped element. If the aspect ratio of K is e.g. h_x/h_y (cf. (4.3) and (4.4)), then that of the corresponding $\mathcal{T}_k$ is $(h_x/h_y)k$.

There is a one-to-one correspondence between Y^h and Y^k given by

$$I^k : Y^h \to Y^k, \quad I^h : Y^k \to Y^h,$$

where I^h is the nodal interpolation operator on Y^h. We use the notation $u_h \in Y^h$ and $u_k \in Y^k$ in order to denote two corresponding functions.

LEMMA 5.1 Let $\hat{K} = (0, h_x) \times (0, h_y) \times (0, h_z)$. Then there exist positive constants c and C, such that, for $u_h \in Y^h(\hat{K})$,

$$\begin{aligned} c\|u_h\|_{0,\hat{K}} &\leqslant \|u_k\|_{0,\hat{K}} \leqslant C\|u_h\|_{0,\hat{K}}, \\ c\|\partial_x(u_h)\|_{0,\hat{K}} &\leqslant \|\partial_x(u_k)\|_{0,\hat{K}} \leqslant C\|\partial_x(u_h)\|_{0,\hat{K}}, \end{aligned}$$

with, in particular, c and C independent of h_x, h_y, h_z, and k. Similar bounds hold for the y and z derivatives. If $K \in \mathcal{T}$ is given by (4.4), then, for $u_h \in Y^h(K)$,

$$\begin{aligned} c\|u_h\|_{0,K} &\leqslant \|u_k\|_{0,K} \leqslant C\|u_h\|_{0,K}, \\ c|u_h|_{1,K} &\leqslant |u_k|_{1,K} \leqslant C|u_h|_{1,K} \end{aligned}$$

where the constants are independent of the diameter and the aspect ratio of K, and k.

The proof of the above result can be found in Canuto (1994, Section 2) for $K = \hat{Q}$. For an affinely mapped element a scaling argument can be used. We note that thanks to Lemma 5.1 we can equivalently work with functions in Y^k or Y^h.

The following result can be found in Casarin (1996, Lemma 3.3.3).

LEMMA 5.2 Let $\hat{K} = (0, h_x) \times (0, h_y) \times (0, h_z)$ and $u_h \in Y^h(\hat{K})$. Given $\theta \in W^{1,\infty}(\hat{K})$, with

$$\|\theta\|_{\infty,\hat{K}} \leqslant C, \quad \|\nabla\theta\|_{\infty,\hat{K}} \leqslant C/r,$$

then

$$\begin{aligned} \|I^h(\theta u_h)\|^2_{0,\hat{K}} &\leqslant C\|u_h\|^2_{0,\hat{K}}, \\ \|\partial_x I^h(\theta u_h)\|^2_{0,\hat{K}} &\leqslant C(|u_h|^2_{1,\hat{K}} + r^{-2}\|u_h\|^2_{0,\hat{K}}), \end{aligned}$$

where C is independent of h_x, h_y, h_z, and k. Similar bounds hold for the y and z derivatives. If $K \in \mathcal{T}$ is given by (4.4), then, for $u_h \in Y^h(K)$,

$$\begin{aligned} \|I^h(\theta u_h)\|^2_{0,K} &\leqslant C\|u_h\|^2_{0,K}, \\ |I^h(\theta u_h)|^2_{1,K} &\leqslant C(|u_h|^2_{1,K} + r^{-2}\|u_h\|^2_{0,K}), \end{aligned}$$

where C is independent of the diameter and the aspect ratio of K, and k.

Given an element $\hat{K} = (0, h_x) \times (0, h_y) \times (0, h_z)$ and a coordinate direction, say x, let a, b, c and d be the vertices of a face of $\hat{K}$ perpendicular to this direction, and let a', b', c' and d' be the corresponding points on the parallel face. The following lemma relies on trivial properties of trilinear functions (cf. Casarin, 1996, Lemma 3.3.1).

LEMMA 5.3 Let $\hat{K} = (0, h_x) \times (0, h_y) \times (0, h_z)$ and a, b, c and d be the vertices of a face of $\hat{K}$ perpendicular to the x direction. Then there are constants independent of h_x, h_y and h_z, such that, if u is trilinear on $\hat{K}$,

$$
\begin{aligned}
c\|u\|^2_{0,\hat{K}} &\leqslant h_x h_y h_z \sum_{\mathbf{x}=a,b,c,d} (u(\mathbf{x})^2 + u(\mathbf{x}')^2) \leqslant C\|u\|^2_{0,\hat{K}},\\
c\|\partial_x u\|^2_{0,\hat{K}} &\leqslant (h_x h_y h_z / h_x^2) \sum_{\mathbf{x}=a,b,c,d} (u(\mathbf{x}) - u(\mathbf{x}'))^2 \leqslant C\|\partial_x u\|^2_{0,\hat{K}},\\
c\|\partial_x u\|^2_{\infty,\hat{K}} &\leqslant h_x^{-2} \sum_{\mathbf{x}=a,b,c,d} (u(\mathbf{x}) - u(\mathbf{x}'))^2 \leqslant C\|\partial_x u\|^2_{\infty,\hat{K}}.
\end{aligned}
$$

Similar bounds hold for the y and z derivatives.

6. Neumann–Neumann methods

Iterative substructuring methods rely on a non-overlapping partition into substructures. We mention Smith *et al.* (1996, Chapter 4) as a general reference to this section. In our algorithms the substructures are chosen as the macroelements in $\mathcal{T}_m = \{\Omega_i \mid 1 \leqslant i \leqslant N\}$. We recall that the macroelements are shape-regular. This appears to be essential for the analysis and good performance.

We define the boundaries $\Gamma_i = \partial\Omega_i \setminus \partial\Omega$ and the interface Γ as their union. We remark that Γ is the union of the interior subdomain *faces*, regarded as open sets, which are shared by two subregions, and subdomain *edges* and *vertices*, which are shared by more than two subregions. Vertices can only be endpoints of edges. In the following, we tacitly assume that points on $\partial\Omega$ are excluded from the geometrical objects that we consider, or, in other words, we will only deal with geometrical objects (faces, edges, vertices, ...) that belong to Γ. We denote the faces of Ω_i by F^{ij}, its edges by E^{ij}, its vertices by V^{ij}, and its *wirebasket*, defined as the union of its edges and vertices, by W^i. Occasionally, we will also use faces, edges and vertices with one or no superscript. If a vertex (edge) lies on $\partial\Omega$ we will regard it as part of the internal edge (resp., face) that shares it with $\partial\Omega$.

When restricted to the subdomain Ω_i, the global triangulation $\mathcal{T}$ determines a local mesh $\mathcal{T}_i$. This mesh can be of four types: face, edge, corner or consisting of just one element. We define the local spaces $X_i = X^k(\Omega_i; \mathcal{T}_i)$, of local finite-element functions that vanish on $\partial\Omega \cap \partial\Omega_i$

In our analysis, we will also employ the GLL mesh $\mathcal{T}_k(\Omega_i)$ on Ω_i, generated by the local GLL meshes $\mathcal{T}_k(K)$ for $K \in \mathcal{T}_i$. The corresponding space of piecewise trilinear functions on $\mathcal{T}_k(\Omega_i)$ that vanish on $\partial\Omega \cap \partial\Omega_i$ is denoted by $Y^h(\Omega_i)$. We set $Y^k(\Omega_i) = X^k(\Omega_i; \mathcal{T}_i)$.

We next define the local bilinear forms

$$
a_i(u, v) = \int_{\Omega_i} \rho_i \nabla u \cdot \nabla v \, d\mathbf{x}, \quad u, v \in X_i.
$$

We note that if Ω_i is a *floating* subdomain (i.e. its boundary does not touch $\partial\Omega$), $a_i(\cdot, \cdot)$ is

only positive semi-definite and for $u \in X_i$ we have

$$a_i(u,u) = 0 \quad \text{iff} \quad u \text{ constant in } \Omega_i.$$

The sets of nodal points on Γ_i, Γ, F^{ij}, E^{ij} and W^i are denoted by $\Gamma_{i,h}$, Γ_h, F_h^{ij}, E_h^{ij} and W_h^i, respectively. We will identify these sets with the corresponding sets of degrees of freedom. As for the corresponding regions, we will also use notation with one or no superscript.

We introduce some spaces defined on the interfaces: U_i is the space of restrictions to Γ_i of functions in $X^k(\Omega_i;\mathcal{T}_i)$ and U of restrictions to Γ of functions in $X^k(\Omega;\mathcal{T})$. We note that functions in U_i and U are uniquely determined by the nodal values in $\Gamma_{i,h}$ and Γ_h, respectively. In the following we will identify these spaces with those of the corresponding harmonic extensions; see in particular Lemma 6.1 below. For every substructure Ω_i, there is a natural interpolation operator

$$R_i^T : U_i \longrightarrow U$$

that extends a function on Γ_i to a global function on Γ with vanishing degrees of freedom in $\Gamma_h \setminus \Gamma_{i,h}$. Its transpose with respect to the Euclidean scalar product $R_i : U \to U_i$ extracts the degrees of freedom in $\Gamma_{i,h}$.

Once a vector $u \in X^k(\Omega;\mathcal{T})$ is expanded using the basis functions introduced in Section 5, problem (3.2) can be written as a linear system

$$Au = f.$$

We recall that the condition number of A is expected to grow at least as $k^3/(h_{min})^2 \sim k^3\sigma^{-2n} \sim k^3\sigma^{-2k}$ (see Melenk, 2002 for a result in two dimensions) and may thus be extremely large for large values of k.

The contributions to the stiffness matrix and the right-hand side can be formed one subdomain at a time. The stiffness matrix is then obtained by *subassembly* of these parts. We will order the nodal points interior to the subdomains first, followed by those on the interface Γ. Similarly, for the stiffness matrix relative to a substructure Ω_i, we have

$$A^{(i)} = \begin{pmatrix} A_{II}^{(i)} & A_{I\Gamma}^{(i)} \\ A_{\Gamma I}^{(i)} & A_{\Gamma\Gamma}^{(i)} \end{pmatrix}. \tag{6.1}$$

In a first step of many iterative substructuring algorithms, the unknowns in the interior of the subdomains are eliminated by block Gaussian elimination. In this step, the Schur complements, with respect to the variables associated with the boundaries of the individual substructures, are calculated. The resulting linear system can be written as

$$Su_\Gamma = g_\Gamma. \tag{6.2}$$

Given the local Schur complements

$$S_i = A_{\Gamma\Gamma}^{(i)} - A_{I\Gamma}^{(i)^T} A_{II}^{(i)^{-1}} A_{I\Gamma}^{(i)} : U_i \longrightarrow U_i,$$

we have

$$S = \sum_{i=1}^{N} R_i^T S_i R_i : U \longrightarrow U$$

and an analogous formula can be found for g_Γ (see Smith *et al.*, 1996, Chapter 4).

A function $u^{(i)}$ defined on Ω_i is said to be discrete harmonic on Ω_i if

$$A_{II}^{(i)} u_I^{(i)} + A_{I\Gamma}^{(i)} u_\Gamma^{(i)} = 0.$$

In this case, it is easy to see that $\mathcal{H}_i(u_\Gamma^{(i)}) := u^{(i)}$ is completely defined by its value on Γ_i. The space of piecewise discrete harmonic functions u consists of functions in X that are discrete harmonic on each substructure. In this case, $u =: \mathcal{H}(u_\Gamma)$ is completely defined by its value on Γ.

Our preconditioners will be defined with respect to the inner product

$$s(u, v) = u^T S v, \quad u, v \in U.$$

It follows immediately from the definition of S that $s(\cdot, \cdot)$ is symmetric and coercive.

The following lemma results from elementary variational arguments.

LEMMA 6.1 Let $u_\Gamma^{(i)}$ be the restriction of a finite-element function to Γ_i. Then the discrete harmonic extension $u^{(i)} = \mathcal{H}_i(u_\Gamma^{(i)})$ of $u_\Gamma^{(i)}$ into Ω_i satisfies

$$a_i(u^{(i)}, u^{(i)}) = \min_{v^{(i)}|_{\partial\Omega_i} = u_\Gamma^{(i)}} a_i(v^{(i)}, v^{(i)}) = {u_\Gamma^{(i)}}^T S^{(i)} u_\Gamma^{(i)}.$$

Analogously, if u_Γ is the restriction of a finite-element function to Γ, the piecewise discrete harmonic extension $u = \mathcal{H}(u_\Gamma)$ of u_Γ into the interior of the subdomains satisfies

$$a(u, u) = \min_{v|_\Gamma = u_\Gamma} a(v, v) = s(u, u) = u_\Gamma^T S u_\Gamma.$$

This lemma ensures that instead of working with functions defined on the interface Γ, we can equivalently work with the corresponding discrete harmonic extensions. For this reason, in the following we will identify spaces of traces on the interfaces, U_i and U, with spaces of discrete harmonic extensions. We point out, however, that due to the particular meshes considered, we cannot equivalently work with norms of local discrete harmonic extensions and traces on the subdomain boundaries since our local meshes are not in general quasi-uniform or shape-regular, and stable discrete harmonic extensions cannot be found in general; see Section 8.

Neumann–Neumann methods provide preconditioners for the Schur complement system: instead of solving (6.2) using, e.g. the conjugate gradient method, they employ an equivalent system involving a preconditioned operator of the form

$$\hat{S}^{-1} S = P_{NN} = P_0 + (I - P_0)\left(\sum_{i=1}^{N} P_i\right)(I - P_0).$$

We refer to Dryja & Widlund (1995), Mandel & Brezina (1996), Pavarino (1997) and Klawonn & Widlund (2001) for some NN methods for the h and p finite-element approximations. We are unaware on any such method for hp-approximations.

The operators P_i are projection-like operators associated to a family of subspaces U_i and determined by a set of local bilinear forms defined on them:

$$\tilde{s}_i(u, v), \quad u, v \in U_i.$$

Given the interpolation operators $R_i^T : U_i \to U$, we have

$$P_i = R_i^T \tilde{P}_i, \quad \tilde{P}_i : U \longrightarrow U_i, \tag{6.3}$$

with

$$\tilde{s}_i(\tilde{P}_i u, v_i) = s(u, R_i^T v_i), \quad v_i \in U_i. \tag{6.4}$$

While P_0 is associated with a low-dimensional global problem, the others are associated with the single substructures. The remainder of this section is devoted to the definition of the various components of P_{NN}.

An important role is played by a family of weighted counting functions δ_i, which are associated with and defined on the individual Γ_i (cf. Dryja *et al.*, 1996; Dryja & Widlund, 1995; Mandel & Brezina, 1996; Sarkis, 1994; Pavarino, 1997) and are defined for $\gamma \in [1/2, \infty)$. Given Ω_i and $\mathbf{x} \in \Gamma_{i,h}$, $\delta_i(\mathbf{x})$ is determined by a sum of contributions from Ω_i and its relevant next neighbours,

$$\delta_i(\mathbf{x}) = \sum_{j \in \mathcal{N}_\mathbf{x}} \rho_j^\gamma(\mathbf{x}) / \rho_i^\gamma(\mathbf{x}), \quad \mathbf{x} \in \Gamma_{i,h}. \tag{6.5}$$

Here $\mathcal{N}_\mathbf{x}$, $\mathbf{x} \in \Gamma_h$, is the set of indices j of the subregions such that $\mathbf{x} \in \Gamma_{j,h}$. These nodal values on $\Gamma_{i,h}$ are then interpolated in order to obtain a function of $\delta_i \in U_i$. The pseudoinverses $\delta_i^\dagger \in U_i$ are defined, for $\mathbf{x} \in \Gamma_{i,h}$, by

$$\delta_i^\dagger(\mathbf{x}) = \delta_i^{-1}(\mathbf{x}), \quad \mathbf{x} \in \Gamma_{i,h}. \tag{6.6}$$

We note that these functions provide a partition of unity:

$$\sum_{i=1}^N R_i^T \delta_i{}^\dagger(\mathbf{x}) \equiv 1. \tag{6.7}$$

In particular, for $u \in U$ we can use the formula

$$u = \sum_{i=1}^N R_i^T u_i, \quad \text{with } u_i = \mathcal{H}_i(\delta_i^\dagger u). \tag{6.8}$$

Here and from now on, we will tacitly assume that whenever we write $\mathcal{H}_i(uv)$ or $\mathcal{H}(uv)$ we first form $I^k(uv)$, i.e. map the product of the two functions u and v into the hp finite-element space by interpolation, and then extend the result as a discrete harmonic function.

If there is no confusion, we will sometimes use the notation uv in order to denote $I^k(uv)$ or $\mathcal{H}_i(uv)$.

A coarse space U_0 of minimal dimension is defined as

$$U_0 = \text{span}\{R_i^T \delta_i^\dagger\} \subset U,$$

where the span is taken over the floating subdomains. We note that U_0 consists of piecewise discrete harmonic functions and R_0^T is the natural injection $U_0 \subset U$. We consider an exact solver on U_0

$$\tilde{s}_0(u, v) := a(\mathcal{H}u, \mathcal{H}v) = a(u, v).$$

For each substructure Ω_i, the local bilinear form is

$$\tilde{s}_i(u, v) := a_i(\mathcal{H}_i(\delta_i u), \mathcal{H}_i(\delta_i v)), \quad u, v \in U_i.$$

For a floating subdomain $\tilde{P}_i$ is defined only for those $u \in U$ for which $s(u, v) = 0$ for all $v = R_i^T v_i$ such that $\mathcal{H}_i(\delta_i v_i)$ is constant on Ω_i. This condition is satisfied if $a(u, R_i^T \delta_i^\dagger) = 0$; we note that $R_i^T \delta_i^\dagger$ is a basis function for U_0. For such subdomains, we make the solution $\tilde{P}_i u$ of (6.4) unique by imposing the constraint

$$\int_{\Omega_i} \mathcal{H}_i(\delta_i \tilde{P}_i u) \mathrm{d}\mathbf{x} = 0, \tag{6.9}$$

which just means that we select the solution orthogonal to the null space of the Neumann operator. Thus, Range($\tilde{P}_i$) has codimension 1 with respect to the space U_i.

We can equally well use matrix notations. Let D_i be the diagonal matrix with the elements $\delta_i^\dagger(\mathbf{x})$ corresponding to the point $\mathbf{x} \in \Gamma_{i,h}$. Then

$$\tilde{s}_i(u, v) = u^T D_i^{-1} S_i D_i^{-1} v.$$

We also have

$$P_i = R_i^T D_i S_i^\dagger D_i R_i S,$$

where $S_i^\dagger$ is a pseudoinverse of S_i. Analogously for the coarse projection,

$$P_0 = R_0^T S_0^{-1} R_0 S,$$

where $S_0 = R_0 S R_0^T$ the restriction of S to U_0

The main result of this paper is a bound for the condition number of P_{NN}. Such bound can be found using the abstract Schwarz theory (see e.g. Smith *et al.*, 1996, Chapter. 6). We refer to Mandel & Brezina (1996), Dryja & Widlund (1995), Pavarino (1997) and Klawonn & Widlund (2001) for similar proofs.

A uniform bound for the smallest eigenvalue can be found using the decomposition (6.8) and the fact that P_0 is an orthogonal projection.

LEMMA 6.2 We have

$$s(P_{NN}u, u) \geqslant s(u, u), \quad u \in U.$$

In order to find a bound for the largest eigenvalue, we need a stability property for the local bilinear forms (see e.g. Smith *et al.*, 1996).

ASSUMPTION 6.1 We have

$$s(R_i^T u_i, R_i^T u_i) \leqslant \omega \tilde{s}_i(u_i, u_i), \qquad u_i \in \text{Range}\,(\tilde{P}_i), \quad i = 1, \ldots, N,$$

with

$$\omega = C\,(1-\sigma)^{-6}\left(1 + \log\left(\frac{k}{1-\sigma}\right)\right)^2$$

and C independent of k, n, σ, γ, the coefficients ρ_i and the diameters H_i.

The proof of Assumption 6.1 is given in Section 9. Assumption 6.1 and a colouring argument provide a bound for the largest eigenvalue (see e.g. Pavarino, 1997, Section 8).

LEMMA 6.3 Let Assumption 6.1 be satisfied. Then

$$s(P_{NN}u, u) \leqslant C\omega s(u, u), \quad u \in U.$$

Consequently, the condition number of P_{NN} satisfies

$$\kappa(P_{NN}) \leqslant C\omega = C\,(1-\sigma)^{-6}\left(1 + \log\left(\frac{k}{1-\sigma}\right)\right)^2.$$

7. Decomposition results

A key ingredient for the proof of Assumption 6.1 and for the analysis of many iterative substructuring methods in three dimensions is a decomposition result for local functions in U_i into face and wirebasket components:

$$u = \sum_j u_{F^{ij}} + u_{W^i}, \quad u \in U_i. \tag{7.1}$$

The face component $u_{F^{ij}}$ vanishes on $\partial\Omega_i \setminus F^{ij}$ and is discrete harmonic. It is uniquely determined by the nodal values in F_h^{ij}. The wirebasket component u_{W^i} is also discrete harmonic and vanishes at all points of $\Gamma_{i,h}$ except at those in W_h^i.

We can further decompose a local function by also defining edge and vertex components:

$$u = \sum_j u_{F^{ij}} + \sum_j u_{E^{ij}} + \sum_j u_{V^{ij}}, \quad u \in U_i, \tag{7.2}$$

where $u_{E^{ij}}$ is discrete harmonic and vanishes on $\partial\Omega_i \setminus E^{ij}$, and $u_{V^{ij}}$ vanishes at all nodes in $\Gamma_{i,h}$ except at the vertex V^{ij}. We recall that we exclude geometrical objects on $\partial\Omega$ and that therefore the sums in (7.1) and (7.2) are taken over faces, edges and vertices that do not belong to $\partial\Omega$. Discrete harmonic functions of type $u_{F^{ij}}$, $u_{E^{ij}}$, $u_{V^{ij}}$ and u_{W^i} are called face, edge, vertex and wirebasket functions, respectively.

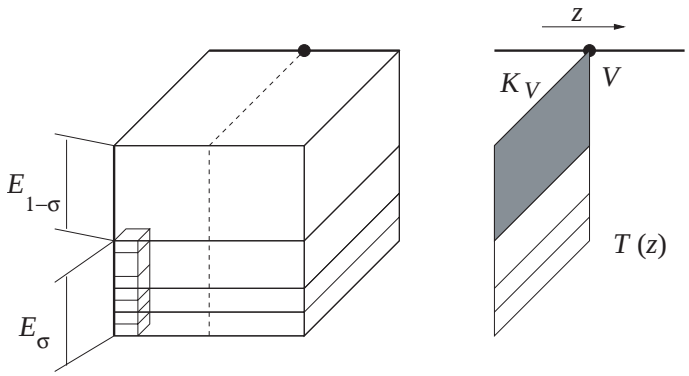


FIG. 2. Face patch: partition of an edge E that touches $\partial\Omega$ into $E_{1-\sigma}$ and E_σ (left) and two-dimensional mesh $\mathcal{T}(z)$ for a section corresponding to a constant z (right).

Here and in the following section, we only carry out proofs for the reference cube $\hat{Q}$: since elements in the macromesh $\mathcal{T}_m$ are shape-regular and affinely mapped, the corresponding bounds for a generic substructure $\Omega_i \in \mathcal{T}_m$, of diameter H_i can be obtained by a standard scaling argument and involve the scaled norm (2.2). We recall that we only need to consider four types of patches: face, edge, corner and unrefined ones, together with the corresponding triangulations $\mathcal{T}_f$, $\mathcal{T}_e$, $\mathcal{T}_c$ and $\hat{Q}$, respectively; cf. Fig. 1. We recall that a generic patch is denoted by Ω_i and its triangulation by $\mathcal{T}_i$.

7.1 *Wirebasket components*

Given an edge $E = E^{ij} \subset W^i$, we define a discrete L^2 norm on E. If E does not touch the boundary $\partial\Omega$, we simply set

$$\|u\|_{h,E} := \|u\|_{0,E}.$$

Let now E be an edge that touches $\partial\Omega$; see Fig. 2, left, for an example of a face patch. After a possible translation and rotation, E can always be written as

$$E = \{(1, 1, z) \mid z \in I\}.$$

Then, the local mesh $\mathcal{T}_i$ gives rise to a one-dimensional triangulation on E, $\mathcal{T}_E$, which is not quasiuniform and is geometrically refined towards one end point, say $z = 1$. In addition, E can be partitioned as

$$\overline{E} = \overline{E}_{1-\sigma} \cup \overline{E}_\sigma, \qquad E_{1-\sigma} = (-1, -1 + 2(1-\sigma)), \quad E_\sigma = (-1 + 2(1-\sigma), 1).$$

We note that $E_{1-\sigma}$ consists of exactly one element of length $2(1-\sigma)$ in $\mathcal{T}_E$, while the elements on E_σ are geometrically refined towards $z = 1$. We now consider the GLL mesh $\mathcal{T}_k(\Omega_i)$ and observe that all the elements that touch the edge E have the same diameters $h_{i,x}$ and $h_{i,y}$, along the two directions perpendicular to E; cf. Fig. 2. Indeed, $h_{i,x}$ and $h_{i,y}$ are of order k^{-2} for a face patch, of order $k^{-2}(1-\sigma)$ for a corner patch and of order k^{-2}

and $k^{-2}(1-\sigma)$, respectively, for an edge patch. Moreover, thanks to our particular meshes and to the fact that local spaces of the same degree k are employed on each element, we have the following property.

PROPERTY 7.1 Let E be an edge parallel to e.g. z, that is shared by two substructures Ω_i and Ω_j. Then, the mesh sizes $h_{i,x}$ and $h_{j,x}$, and $h_{i,y}$ and $h_{j,y}$ are comparable. In particular, there exist constants, depending only on the aspect ratios of Ω_i and Ω_j, such that

$$c(1-\sigma)h_{i,x} \leqslant h_{j,x} \leqslant C(1-\sigma)^{-1}h_{i,x}.$$

Similar bounds hold for $h_{i,y}$ and $h_{j,y}$.

We define

$$\|u\|^2_{h,E} := \|u\|^2_{0,E} + \|u\|^2_{h,E_\sigma} = \|u\|^2_{0,E} + h_{i,x}h_{i,y}\|\partial_z u\|^2_{0,E_\sigma}.$$

We note that in this case the discrete norm is obtained by adding to the L^2 norm on E a weighted L^2 norm of $\partial_z u$ over a part of E where $\mathcal{T}_E$ is not quasiuniform. A discrete wirebasket norm is obtained by summing the contributions over all the edges:

$$\|u\|^2_{h,W^i} := \sum_{E\subset W^i} \|u\|^2_{h,E}.$$

LEMMA 7.1 Let $u_{W^i} \in U_i$ be discrete harmonic and vanish at all nodal points $\Gamma_{i,h}$ except at those on W^i. Then there is a constant independence of u_{W^i}, H_i, σ and n, such that

$$|u_{W^i}|^2_{1,\Omega_i} \leqslant C(1-\sigma)^{-2}\|u_{W^i}\|^2_{h,W^i}.$$

Proof. The result follows by estimating the energy norm of the zero extension of the boundary values and by noting that the harmonic extension has a smaller energy (cf. Lemma 6.1). More precisely, let u_k be the function that vanishes at all nodal points in $\Omega_{i,h} \cup \Gamma_{i,h}$ except at those on W^i, and $u = u_h = I^h u_k$ the corresponding piecewise trilinear function defined on the GLL mesh $\mathcal{T}_k(\Omega_i)$. We will estimate the energy of u_h on each element $K \in \mathcal{T}_k(\Omega_i)$ that touch an edge $E \subset W^i$. Without loss of generality, we assume that E is parallel to the z axis. We only consider the worst possible case, i.e. that of a face patch and refer to Fig. 2.

Let us first suppose that E does not touch $\partial\Omega$. For a face patch, K has dimensions h_x, h_y and h_z of order

$$k^{-2} \times k^{-2}(1-\sigma) \times k^{-2},$$

or

$$k^{-2} \times k^{-2}(1-\sigma) \times k^{-1},$$

and thus

$$\begin{aligned} &c(1-\sigma)h_x \leqslant h_y \leqslant Ch_x, \\ &h_x \leqslant Ch_z; \end{aligned} \tag{7.3}$$

see Fig. 2. If a and b are the vertices of K that lie on E, Lemma 5.3 yields

$$\|\partial_x u\|^2_{0,K} \leqslant C(h_y h_z/h_x)\,(u(a)^2 + u(b)^2) \leqslant C\int_a^b u^2 \mathrm{d}z,$$

where for the last inequality we have used (7.3) and standard properties of linear functions. In a similar way, we find

$$\|\partial_y u\|^2_{0,K} \leqslant C(1-\sigma)^{-1}\int_a^b u^2 \mathrm{d}z, \qquad \|\partial_z u\|^2_{0,K} \leqslant C\int_a^b u^2 \mathrm{d}z.$$

Let now E be an edge that touches $\partial\Omega$ and $K \in \mathcal{T}_k(\Omega_i)$ be an element that shares an edge with $E_{1-\sigma}$. For a face patch, K has dimensions of the order

$$k^{-2} \times k^{-2} \times k^{-2}(1-\sigma),$$

or

$$k^{-2} \times k^{-2} \times k^{-1}(1-\sigma),$$

and thus

$$\begin{aligned} &ch_x \leqslant h_y \leqslant Ch_x, \\ &h_x \leqslant C(1-\sigma)^{-1}h_z; \end{aligned} \tag{7.4}$$

see Fig. 2, left. As before, Lemma 5.3 yields

$$\|\partial_x u\|^2_{0,K} \leqslant C\int_a^b u^2 \mathrm{d}z,\ \|\partial_y u\|^2_{0,K} \leqslant C\int_a^b u^2 \mathrm{d}z,\ \|\partial_z u\|^2_{0,K} \leqslant C(1-\sigma)^{-2}\int_a^b u^2 \mathrm{d}z.$$

We are now left with the case of an element $K \in \mathcal{T}_k(\Omega_i)$ that shares an edge with E_σ. We note that the first of (7.4) remains valid in this case. We then have

$$\|\partial_x u\|^2_{0,K} \leqslant C\int_a^b u^2 \mathrm{d}z, \qquad \|\partial_y u\|^2_{0,K} \leqslant C\int_a^b u^2 \mathrm{d}z.$$

For $\partial_z u$, we trivially have

$$\|\partial_z u\|^2_{0,K} \leqslant C(h_x h_y/h_z)(u(a)-u(b))^2 \leqslant Ch_x h_y \int_a^b (\partial_z u)^2 \mathrm{d}z.$$

The proof is concluded by summing over the elements $K \in \mathcal{T}_k(\Omega_i)$ and using Lemma 5.1. □

We now have a bound for the wirebasket component.

COROLLARY 7.2 Let $u \in U_i$ and u_{W^i} be its wirebasket component. Then there is a constant independent of u, H_i, σ and n such that

$$|u_{W^i}|^2_{1,\Omega_i} \leqslant C(1-\sigma)^{-2}\|u\|^2_{h,W^i}.$$

A complementary result is given by the trace estimates in Lemma 7.3. We first introduce some additional notation. Let E be an edge of a substructure Ω_i. Without loss of generality, we assume that Ω_i coincides with the reference cube $\hat{Q}$ and that $E = \{(1, 1, z) \mid z \in I\}$. The intersection between the plane corresponding to a constant $z \in I$ and $\hat{Q}$ is the unit square $\hat{S} = (-1, 1)^2$, and the local mesh $\mathcal{T}_i$ gives rise to a two-dimensional mesh $\mathcal{T}(z)$ on $\hat{S}$ which is either a two-dimensional edge or corner patch, or it consists of a single element $\hat{S}$; see Fig. 2, right. Let $V = (1, 1)$ be the intersection between E and the closure of $\hat{S}$. If $K_V \in \mathcal{T}(z)$ is the two-dimensional element that contains V, we note that, since E does not belong to $\partial\Omega$, K_V has dimensions in $\{2, 2(1-\sigma)\}$, and thus is independent of the level of refinement n. For a fixed $(x, y) \in \overline{K}_V$, we finally define the edge $E(x, y) = \{(x, y, z) \mid z \in I\}$.

LEMMA 7.3 Let $u_k \in X_i$ and E and edge of Ω_i. Then there is a constant independent of u_k, H_i, σ and n such that

$$\begin{aligned} \|u_k\|^2_{0,E} &\leqslant C\,(1-\sigma)^{-2}\,(1+\log k)\,\|u_k\|^2_{1,\Omega_i}, \\ \|u_k\|^2_{h,E} &\leqslant C\,(1-\sigma)^{-2}\,(1+\log k)\,\|u_k\|^2_{1,\Omega_i}. \end{aligned}$$

Proof. As before, it is enough to find bounds for $u = I^h u_k$. Without loss of generality, we assume $E = \{(1, 1, z) \mid z \in I\}$. We consider the two-dimensional mesh $\mathcal{T}(z)$ on the intersection between the plane corresponding to a constant z and the substructure; cf. Fig. 2, right. Since geometric refinement on $\mathcal{T}(z)$ takes place far from the vertex $(1, 1)$, we can apply the two-dimensional result in Toselli & Vasseur (2003a, Lemma 7.6) and write

$$|u(1, 1, z)|^2 \leqslant C\,(1-\sigma)^{-2}\,(1+\log k)\,\|u(\cdot, \cdot, z)\|^2_{1,\hat{S}}, \quad z \in (-1, 1),$$

with a constant that is independent of n, σ and z. Integrating over z then gives

$$\|u\|^2_{0,E} \leqslant C\,(1-\sigma)^{-2}\,(1+\log k)\,\|u\|^2_{1,\Omega_i},$$

which proves the first inequality and the second one for edges that do not touch $\partial\Omega$.

We now bound $\|u\|_{h,E_\sigma}$ for an edge that touches the boundary $\partial\Omega$. We consider the one-dimensional GLL meshes for each one of the elements in $\mathcal{T}_E$ and estimate the single contributions from the elements of these meshes. Let e be one of these elements of length h_z and end points a and b. The edge e belongs to a parallelepiped $K_e \in \mathcal{T}_k(\Omega_i)$. We note that K_e has dimensions $h_x = h_{i,x}$, $h_y = h_{i,y}$, and h_z. Since u is linear on e and trilinear on K_e, we have

$$h_x h_y \int_e \partial_z u^2 \mathrm{d}z \leqslant C\frac{h_x h_y}{h_z}(u(a) - u(b))^2 \leqslant C\|\partial_z u\|^2_{0,K_e},$$

where, for the last inequality, we have used Lemma 5.3. Summing over the edges e in E_σ yields

$$\|u\|^2_{h,E_\sigma} \leqslant C\|\partial_z u\|^2_{0,\Omega_i},$$

which, combined with the first inequality, proves the second bound. □

The next lemma can be proved using the two-dimensional bound in Toselli & Vasseur (2003a, Lemma 7.6) and similar arguments as before. We note that it is only valid for edges $E(x, y)$ that are not too far from E and thus not too close to the part of Ω_i where anisotropic refinement takes place.

LEMMA 7.4 Let E be an edge of a substructure Ω_i which is parallel, say, to z and intersects the plane corresponding to a constant z in V. Let in addition K_V be the element in the two-dimensional mesh $\mathcal{T}(z)$ that contains V. Then, for every $(x, y) \in \overline{K}_V$ and $u_k \in X_i$,

$$\|u_k\|^2_{0,E(x,y)} \leqslant C\,(1-\sigma)^{-2}\,(1+\log k)\,\|u_k\|^2_{1,\Omega_i}, \tag{7.5}$$

where C is independent of u_k, σ, n, k, and (x, y), but depends only on the aspect ratio of Ω_i.

Proof. The proof can be carried out as in the previous lemma by using the two-dimensional result in Toselli & Vasseur (2003a, Lemma 7.6). Indeed, since the point (x, y) belongs to $\overline{K}_V$ and is thus far from the region where anisotropic refinement takes place, we have

$$|u(x, y, z)|^2 \leqslant C\,(1-\sigma)^{-2}\,(1+\log k)\,\|u(\cdot,\cdot,z)\|^2_{1,\hat{S}}, \quad z \in (-1, 1).$$

Integration along z concludes the proof. □

We end this section with a stability result for vertex and edge components. It is a direct consequence of (5.1) and of the fact that for a vertex function the modified norm $\|\cdot\|_{h,E}$ coincides with $\|\cdot\|_{0,E}$.

LEMMA 7.5 Let E be an edge of a substructure Ω_i and V one of its end points. Then, for every $u \in X_i$,

$$\|u_V\|^2_{h,W^i} \leqslant C\|u\|^2_{h,W^i}, \quad \|u_E\|^2_{h,W^i} \leqslant C\|u\|^2_{h,W^i}, \tag{7.6}$$

where C is independent of u, σ, n, k.

7.2 *Face components*

We next consider the face contributions of the decomposition (7.1). Bounds for face contributions on the unrefined patch follow from standard results for spectral elements. For face, edge and corner patches, we employ cut-off functions θ_F for each face and Lemma 5.2. We note that we need to consider one possible case for faces of the corner patch, and two for the edge and face patches; cf. Fig. 1. In this section we only consider the case of an edge patch Ω_i in full detail, with the edge $(1, y, -1)$, $y \in I$, and the two adjacent faces in common with $\partial\Omega$; see Fig. 3. The other patches can be dealt with in a similar way.

As shown in Fig. 3 for the reference cube, the edges that do not lie on $\partial\Omega$ are denoted by E^l, $l = 1, \ldots, 5$, with E^5 the edge that does not touch the boundary $\partial\Omega$. An edge patch is further partitioned into three regions. The first step of geometric refinement partitions $\hat{Q}$ into four parallelepipeds with dimensions in $\{2, 2(1-\sigma), 2\sigma\}$. Let K_Ω be the one that

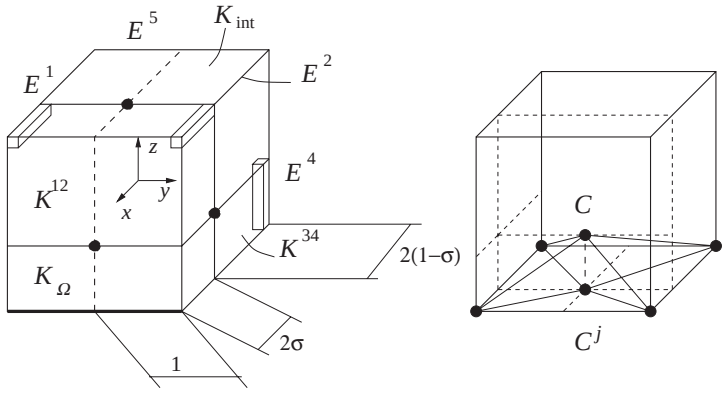


FIG. 3. Edge patch on the reference cube $(-1, 1)^3$ employed in the proofs of Lemmas 7.6 and 7.7.

contains the boundary edge and K_{int} the one that does not touch $\partial\Omega$ and contains the inner edge E^5. The two remaining parallelepipeds are denoted by K^{12} and K^{34} and they touch the edges E^1 and E^2, and E^3 and E^4, respectively. The region K_{edge} is the union of K^{12} and K^{34}; cf. Fig. 3.

The proof of the following lemma is a modification of those of Casarin (1996, Lemma 3.3.6) and Toselli & Vasseur (2003a, Lemma 7.7).

LEMMA 7.6 Given a face F^j of Ω_i that does not lie on $\partial\Omega$, there exists a continuous function θ_{F^j}, defined on $\overline{\Omega_i}$, that is equal to one at the nodal points of F_h^j and zero on $\Gamma_{i,h} \setminus F_h^j$, such that

$$\begin{aligned}
&\sum_{F^j \subset \Gamma_i} \theta_{F^j}(\mathbf{x}) = 1, \quad \mathbf{x} \in (\Omega_{i,h} \cup \Gamma_{i,h}) \setminus W_h^i, \\
&0 \leqslant \theta_{F^j} \leqslant 1, \\
&|\nabla \theta_{F^j}| \leqslant C/r, \quad \text{in } \Omega_i \setminus K_\Omega \\
&|\nabla \theta_{F^j}| \leqslant C/H_i, \quad \text{in } K_\Omega,
\end{aligned} \tag{7.7}$$

where $r = r(\mathbf{x})$ is the distance to the closest edge of Ω_i that does not lie on $\partial\Omega$.

Proof. We only need to construct four functions and we will do that by constructing them in the three regions K_{int}, K_{edge}, and K_Ω separately.

We start with the inner region K_{int} and employ a similar construction as in Casarin (1996, Lemma 3.3.6). We further partition Ω_i into eight parallelepipeds by bisecting $\{K_{int}, K^{12}, K^{34}, K_\Omega\}$ with the plane $y = 0$; see Fig. 3, left. Let the centre C be the common vertex to these parallelepipeds and $\{C^j, \ j = 1, \ldots, 6\}$ be their vertices that belong to the six faces of Ω_i; see Fig. 3, right. By connecting the centre C with the centres C^j and with the eight vertices of Ω_i, and, for each face, by connecting the point C^j with the four vertices of this face, we can partition Ω_i into 24 tetrahedra; see Fig. 3, right. By intersecting them with K_{int}, we obtain a partition of K_{int} into eight tetrahedra. We first define a function ϑ_{F^j} associated with the face F^j, defined to be $1/4$ at the centre C

and $\vartheta_{Fj}(C^l) = \delta_{jl}$ at the centres of the faces. On the segments CC^l, these functions are obtained by linear interpolation of the values at C and C^l; see Fig. 3, right. The values inside each subtetrahedron formed by the segment CC^l and one edge of F^l are defined to be constant on the intersection of any plane through that edge, and are given by the value on the segment CC^l. We note that this procedure determines ϑ_{Fj} at all points in Ω_i except on the wirebasket W^i.

We next consider the GLL triangulation $\mathcal{T}_k(\Omega_i)$ and interpolate ϑ_{Fj} at the GLL nodes in $\overline{K}_{\text{int}} \setminus W^i$:

$$\theta_{Fj}(\mathbf{x}) = (I^h \vartheta_{Fj})(\mathbf{x}), \qquad \mathbf{x} \in \overline{K}_{int} \setminus W^i.$$

The function θ_{Fj} is set to zero on the nodes in W_h^i. The functions θ_{Fj} are non-negative and bounded by one: this proves the second of (7.7) for points in K_{int}. By construction, also the first of (7.7) holds for every node in $\overline{K}_{int} \setminus W^i$. The third of (7.7) can be proven by proceeding in the same way as for Casarin (1996, Lemma 3.3.6).

We next construct the functions θ_{Fj} in K_{edge}. We start with K^{12}. We take the values on the common face $\overline{K^{12}} \cap \overline{K}_{int}$ and we extend them as constants into $\overline{K^{12}}$ along the segments parallel to E^1 and E^2; see Fig. 3, left. The inequalities in (7.7) remain valid. We note that the function obtained is independent of x in K^{12}. A similar construction is carried out in K^{34}.

Finally, we construct θ_{Fj} in K_Ω. We note that K_Ω is divided into two parallelepipeds and that on their internal faces the function θ_{Fj} has already been defined. In addition, θ_{Fj} is bilinear on these faces. It is then enough to assign the value $1/4$ at the end points and mid-point of the boundary edge and interpolate these values in K_Ω in order to obtain a piecewise trilinear function. The first, second and fourth of (7.7) follow from standard properties of trilinear functions. □

By examining the proof of the previous lemma, we see that, for an edge E that touches $\partial\Omega$, the value of the functions θ_{Fj} is independent of the coordinate along the direction of E in all the elements of the GLL meshes that touch E_σ; cf. Fig. 3, left.

PROPERTY 7.2 Let F be a face of Ω_i and E be an edge, parallel to say z, that touches $\partial\Omega$. In any element $K_E \in \mathcal{T}_k(\Omega_i)$ that shares an edge with E_σ the function θ_F is independent of z.

We are now able to bound the face components in the decomposition (7.1).

LEMMA 7.7 Let θ_{Fj} be the functions in Lemma 7.6, where F^j is a face of the substructure Ω_i. Then, for every $\mathbf{x} \in \Omega_{i,h} \cup \Gamma_{i,h}$ that is not on the wirebasket of Ω_i,

$$\sum_j I^k(\theta_{Fj} u)(\mathbf{x}) = \sum_j I^h(\theta_{Fj} u)(\mathbf{x}) = u(\mathbf{x}), \quad u \in X_i$$

and

$$|I^k(\theta_{Fj} u)|^2_{1,\Omega_i} \leqslant C\,(1-\sigma)^{-4} \left(1 + \log\left(\frac{k}{1-\sigma}\right)\right)^2 ||u||^2_{1,\Omega_i}.$$

Proof. We only consider the case of an edge patch Ω_i in full detail; see Fig. 3. The proof is similar to that in Toselli & Vasseur (2003a, Lemma 7.8) and Casarin (1996, Lemma 3.3.7) but particular care is required close to the edges that touch $\partial\Omega$. Indeed, thanks to Lemma 5.1, it is enough to find a bound for the piecewise trilinear function $I^h(\theta_{Fj}u)$.

The first equality follows directly from the first of (7.7). For the second inequality, we consider an element K, of dimensions h_x, h_y, and h_z, in the GLL mesh $\mathcal{T}_k(\Omega_i)$. We consider three cases (as opposed to Casarin, 1996, Lemma 3.3.7 where only two cases are considered): K may belong to the region K_Ω containing the boundary edge, touch the wirebasket, or may not touch it; see Fig. 3.

Case 1. We start with an element that touches an edge E and does not belong to K_Ω. We can proceed as in Casarin (1996, Lemma 3.3.7) if E does not touch $\partial\Omega$ ($E = E^5$) or, in case it does ($E = E^l$, $l = 1, \ldots, 4$), if K does not touch E_σ. We only consider the case of $E = E^3$ in full detail; cf. Fig. 3, left. The nodal values of $I^h(\theta_{Fj}u)$ on K are 0, 0, 0, 0, $u(a)$, $u(b)$, $\theta_{Fj}(c)u(c)$ and $\theta_{Fj}(d)u(d)$, with a and b vertices on a face and c and d vertices inside Ω_i. It is immediate to see that

$$\begin{aligned} &c(1-\sigma)h_x \leqslant h_y \leqslant C(1-\sigma)^{-1}h_x, \\ &h_x \leqslant C(1-\sigma)^{-1}h_z. \end{aligned} \tag{7.8}$$

Using Lemma 5.3 and (7.8), we can easily find

$$\begin{aligned} |I^h(\theta_{Fj}u|^2_{1,K} &\leqslant C(1-\sigma)^{-2}h_z\,(u(a)^2 + u(b)^2 + u(c)^2 + u(d)^2) \\ &\leqslant C(1-\sigma)^{-2}\left(\int_a^b u^2 \mathrm{d}z + \int_c^d u^2 \mathrm{d}z\right), \end{aligned}$$

where we have also used the fact that θ_{Fj} has values between zero and one. Summing over the element K and using in Lemma 7.4 for segments that are parallel to E gives

$$\sum_K |I^h(\theta_{Fj}u)|^2_{1,K} \leqslant C\,(1-\sigma)^{-4}\,(1+\log k)\,\|u\|^2_{1,\Omega_i},$$

where the sum is taken over the elements in $\mathcal{T}_k(\Omega_i)$ that touch an edge E, such that E does not touch $\partial\Omega$ or, if it does, K does not touch E_σ.

We next consider the case where K shares an edge with E_σ. The terms involving the x and y derivatives can be bounded as before: indeed, the first of (7.8) still holds in this case. However, the second of (7.8), needed to bound the z derivative, does not hold. Using Lemma 5.3 we find

$$\|\partial_z I^h(\theta_{Fj}u)\|^2_{0,K} \leqslant C(h_x h_y/h_z)\left((u(a)-u(b))^2 + (\theta_{Fj}(c)u(d) - \theta_{Fj}(d)u(d))^2\right).$$

Property 7.1 ensures that $\theta_{Fj}(c) = \theta_{Fj}(d)$ and thus

$$\|\partial_z I^h(\theta_{Fj}u)\|^2_{0,K} \leqslant C\|\partial_z(\theta_{Fj}u)\|^2_{0,K}.$$

Summing over the elements K that touch E_σ gives

$$\sum_K \|\partial_z(I^h(\theta_{Fj}u)\|^2_{0,K} \leqslant C\,\|\partial_z(\theta_{Fj}u)\|^2_{0,\Omega_i}$$

and thus

$$\sum_{\overline{K}\cap W^i\neq\emptyset} |I^h(\theta_{Fj}u)|^2_{1,K} \leqslant C\,(1-\sigma)^{-4}\,(1+\log k)\,\|u\|^2_{1,\Omega_i}. \tag{7.9}$$

Case 2. We now consider an element $K \in \mathcal{T}_k(\Omega_i)$ that does not touch the wirebasket and does not belong to K_Ω. The proof for this case is similar to that of Casarin (1996, Lemma 3.3.7). Using Lemma 5.2 and the second of (7.7), we have

$$\sum_{\substack{K\subset\Omega_i\setminus K_\Omega\\ \overline{K}\cap W^i=\emptyset}} |I^h(\theta_{Fj}u)|^2_{1,K} \leqslant C\sum_K (|u|^2_{1,K} + r_K^{-2}\|u\|^2_{0,K}),$$

where r_K is the distance of the baricentre of K from the wirebasket. We have

$$\begin{aligned}\sum_K r_K^{-2}\|u\|^2_{0,K} &\leqslant C \int\limits_{K_{int}\cup K^{12}\cup K^{34}} r^{-2}u^2\mathrm{d}\mathbf{x}\\ &\leqslant C\int\limits_{K_{int}} r_5^{-2}u^2\mathrm{d}\mathbf{x} + C\sum_{j=1}^{2}\int\limits_{K^{12}\cup K_{int}} r_j^{-2}u^2\mathrm{d}\mathbf{x} + C\sum_{j=3}^{4}\int\limits_{K^{34}\cup K_{int}} r_j^{-2}u^2\mathrm{d}\mathbf{x},\end{aligned}$$

where r_j denotes the distance of a point from the edge E^j, and the region consisting of the elements in the GLL mesh $\mathcal{T}_k(\Omega_i)$ that touch the wirebasket is assumed to be excluded from the domains of integration; cf. Fig. 3, left. Each of the integrals on the right, associated with an edge $E = E^j$, can be estimated using cylindrical coordinates with the ζ axis coinciding with E^j and the radial direction r_j normal to E^j. We only consider E^5 in detail; cf. Fig. 3. The other integrals can be estimated in the same way. If the point V is the intersection between E^5 and the section corresponding to a fixed ζ, and K_V is the element of the two-dimensional mesh $\mathcal{T}(\zeta)$ that contains V, we can write

$$\begin{aligned}\int\limits_{K_{int}} r_5^{-2}u^2\mathrm{d}\mathbf{x} &\leqslant C\int\limits_{K_V} r_5^{-2}\,\mathrm{d}x\mathrm{d}y\int\limits_{-1}^{1} u^2\mathrm{d}\zeta\\ &\leqslant C(1-\sigma)^{-2}(1+\log k)\|u\|^2_{1,\Omega_i}\int\limits_{K_V} r_5^{-2}\,\mathrm{d}x\mathrm{d}y,\end{aligned}$$

where we have used Lemma 7.4 for the last inequality; cf. Fig. 2, right. The last integral can be estimated by

$$\int\limits_{K_V} r_5^{-2}\,\mathrm{d}x\mathrm{d}y \leqslant C\int_{k^{-2}(1-\sigma)}^{2} r_5^{-1}\,\mathrm{d}r_5\int_0^{2\pi}\mathrm{d}\phi \leqslant C\left(1+\log\left(\frac{k}{1-\sigma}\right)\right).$$

Considering similar contributions for the other edges, we then find

$$\sum_{\substack{K\subset\Omega_i\setminus K_\Omega\\ \overline{K}\cap W^i=\emptyset}} |I^h(\theta_{Fj}u)|^2_{1,K} \leqslant C|u|^2_{1,\Omega_i} + C(1-\sigma)^{-2}\left(1+\log\left(\frac{k}{1-\sigma}\right)\right)^2\|u\|^2_{1,\Omega_i}. \tag{7.10}$$

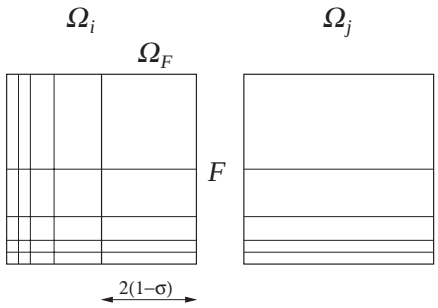


FIG. 4. The cross sections of an edge and a face patch, or a corner and an edge patch, with a common face F.

Case 3. We are now left with the case $K \subset K_\Omega$. Since, in this case, $|\nabla\theta_{Fj}|$ is bounded by a constant, Lemma 5.2 ensures

$$\sum_{K\subset K_\Omega} |I^h(\theta_{Fj}u)|^2_{1,K} \leqslant C\,\|u\|^2_{1,\Omega_i}.$$

The proof is concluded by combining this inequality with (7.9) and (7.10), and applying Lemma 5.1. □

8. Comparison results

In the analysis of many iterative substructuring methods, it is necessary to compare certain norms of discrete harmonic functions on different substructures that have the same trace on a common face, edge or vertex.

As already pointed out in Toselli & Vasseur (2003a), if the local meshes are shape-regular and quasi-uniform, the comparison for functions on adjacent substructures that have the same value on a common face can be made using a trace theorem (which is valid for general functions in H^1) and a stable extension from the face. However, the existence of stable extensions for meshes that are not quasi-uniform or shape-regular is far from trivial. For this reason, here we will adopt the same strategy as in Toselli & Vasseur (2003a), since the meshes considered are highly anisotropic but of a particular type.

We note that we only need to consider three cases: that of a face shared by an unrefined and a face patch, by a face and an edge patch, and by an edge and a corner patch. We only consider the last two cases in full detail, since the former can be treated in exactly the same way. We consider the two substructures Ω_i and Ω_j in Fig. 4, which share the face F. Since we proceed in exactly the same way as in Toselli & Vasseur (2003a, Section 7.3), we do not present any proof here. We first consider Ω_i and suppose that it coincides with the reference cube $\hat{Q}$. The face F corresponds to $x = 1$. Let Ω_F be the layer of points in Ω_i within a distance $2(1-\sigma)$ from F.

The following lemma can be proven in the same way as Toselli & Vasseur (2003a, Lemma 7.9).

LEMMA 8.1 Let $u_F \in U_i$ be a face function on Ω_i, i.e. a discrete harmonic function that vanishes on $\partial\Omega_i \setminus F$, and $\tilde{u}_F \in X_i$, such that

1. $\tilde{u}_F$ is equal to u_F on F and vanishes on $\partial\Omega_F \setminus F$;
2. $\tilde{u}_F$ is discrete harmonic in Ω_F;
3. $\tilde{u}_F$ vanishes in $\Omega_i \setminus \Omega_F$.

Then

$$|u_F|^2_{1,\Omega_i} \leqslant |\tilde{u}_F|^2_{1,\Omega_i} \leqslant \|\nabla\theta_{\sigma,F}\|^2_\infty \, |u_F|^2_{1,\Omega_i},$$

where $\theta_{\sigma,F} \in W^{1,\infty}(\Omega_i)$ is any function that is equal to one on F, vanishes in $\Omega_i \setminus \Omega_F$, and has values in $(0, 1)$ in the rest of Ω_i. In particular, we can find a function such that

$$\|\nabla\theta_{\sigma,F}\|_\infty \leqslant C(1-\sigma)^{-1}.$$

The comparison result for face functions can be then found by noting that we can map Ω_j and its mesh into Ω_F and the corresponding local mesh, by a simple dilation in the horizontal direction.

COROLLARY 8.2 Let F be a face that is common to Ω_i and Ω_j and $u_F \in U$ be a piecewise discrete harmonic function that is identically zero at all nodal points in $\Gamma_h \setminus F_h$. Then,

$$c\,(1-\sigma)\,|u_F|^2_{1,\Omega_i} \leqslant |u_F|^2_{1,\Omega_j} \leqslant C(1-\sigma)^{-1}\,|u_F|^2_{1,\Omega_i}.$$

For vertex and edge functions the following lemma is sufficient for our analysis.

LEMMA 8.3 Let Ω_i and Ω_j be two substructures and $u \in X$. If $V = V^i = V^j$ is a common vertex, then the vertex components of u satisfy

$$\|u_{V^j}\|^2_{h,W^j} \leqslant C(1-\sigma)^{-1}\|u_{V^i}\|^2_{h,W^i}.$$

If $E = E^i = E^j$ is a common edge, then the edge components of u satisfy

$$\|u_{E^j}\|^2_{h,W^j} \leqslant C(1-\sigma)^{-2}\|u_{E^i}\|^2_{h,W^i}.$$

Proof. For the first inequality, we note that the modified norms $\|\cdot\|_{h,W^i}$ and $\|\cdot\|_{h,W^j}$ coincide with the L^2 norms, since a vertex function vanishes at all nodal points in Γ_h except at that vertex and we only consider internal vertices. It is enough to compare a contribution from an edge E^j of Ω_j with that of an edge E^i of Ω_i. The worst possible case occurs when E^j does not touch $\partial\Omega$ but E^i does; cf. Fig. 4. Let $\phi(\hat{z})$ be the function in $\mathbb{Q}_k(I)$ that vanishes at all the GLL nodes in I, except at -1 where it is equal to $u(V)$. Using the change of variables $z = (1-\sigma)(\hat{z}+1) - 1$ and the fact that u_{V^i} vanishes in E^i_σ, we have

$$\begin{aligned}\int_{E^j} u_{V^j}(\hat{z})^2 \mathrm{d}\hat{z} &= \int_{-1}^{1} \phi(\hat{z})^2 \mathrm{d}\hat{z} = (1-\sigma)^{-1}\int_{-1}^{-1+2(1-\sigma)} \phi(z)^2 \mathrm{d}z \\ &= (1-\sigma)^{-1}\int_{E^i_{1-\sigma}} u_{V^i}(z)^2 \mathrm{d}z = (1-\sigma)^{-1}\int_{E^i} u_{V^i}(z)^2 \mathrm{d}z.\end{aligned}$$

For the second inequality, it is enough to use the definition of the modified norms $\|\cdot\|_{h,W^i}$ and $\|\cdot\|_{h,W^j}$ and Property 7.1 □

9. Proof of Assumption 6.1

We are now ready to give an upper bound for ω in Assumption 6.1. Our proof is similar to that in Pavarino (1998, Lemma 9.1). We note that if $u_i \in U_i$, its extension $u = R_i^T u_i$ vanishes on Γ_h except at the nodal points in $\Gamma_{i,h}$ and its support is thus contained in the union of Ω_i and its neighbouring substructures. In order to estimate ω we thus have to estimate the energy of u in these substructures in terms of the energy of $\mathcal{H}_i(\delta_i u_i)$ in Ω_i alone.

We first note that, by simple calculation, we have

$$\rho_j(\delta_i^\dagger(\mathbf{x}))^2 = \rho_j \delta_i(\mathbf{x})^{-2} \leqslant \min\{\rho_i, \rho_j\}, \qquad \mathbf{x} \in \Gamma_{i,h}, \quad j \in \mathcal{N}_\mathbf{x}. \tag{9.1}$$

Let $u_i \in \text{Range}(\tilde{P}_i)$. We start with a substructure Ω_j that only has a vertex $V = V^i = V^j$ in common with Ω_i. We note that, according to the decomposition (7.2), u has only a wirebasket component $u_{Vj} = u$ on Ω_j, which vanishes at all nodes in $\Gamma_{j,h}$ except at V. Using Lemma 7.1, we find

$$\begin{aligned} a_j(u,u) = \rho_j |u_{Vj}|^2_{1,\Omega_j} &\leqslant C\, \rho_j\, (1-\sigma)^{-2}\, \|u_{Vj}\|^2_{h,W^j} \\ &= C\, \rho_j\, \delta_{i,V}^{-2}\, (1-\sigma)^{-2}\, \|\delta_i u_{Vj}\|^2_{h,W^j}, \end{aligned}$$

where $\delta_{i,V} = \delta_i(V)$. We next note that, thanks to Lemma 8.3, the norm $\|\cdot\|_{h,W^j}$ associated with Ω_j can be bounded by $\|\cdot\|_{h,W^i}$. In addition, we can apply Lemmas 7.5 and 7.3 and find

$$\begin{aligned} \rho_i \|\delta_i u_{Vj}\|^2_{h,W^j} &\leqslant C(1-\sigma)^{-1}\rho_i \|(\delta_i u_i)_{V^i}\|^2_{h,W^i} \leqslant C(1-\sigma)^{-1}\rho_i \|\mathcal{H}_i(\delta_i u_i)\|^2_{h,W^i} \\ &\leqslant C(1-\sigma)^{-3}(1+\log k)\ \rho_i \|\mathcal{H}_i(\delta_i u_i)\|^2_{1,\Omega_i} \\ &= C(1-\sigma)^{-3}(1+\log k)\ (a_i(\mathcal{H}_i(\delta_i u_i), \mathcal{H}_i(\delta_i u_i)) + \rho_i H_i^{-2} \|\mathcal{H}_i(\delta_i u_i)\|^2_{0,\Omega_i}). \end{aligned}$$

The L^2 component in the last term can be bounded by the local bilinear form $a_i(\cdot,\cdot)$, thanks to a Poincaré inequality for floating subdomains (cf. (6.9)), or thanks to a Friedrichs inequality for substructures that touch $\partial\Omega$. Combining these two estimates and using (9.1), we find

$$a_j(u,u) = a_j(u_{Vj}, u_{Vj}) \leqslant C(1-\sigma)^{-5}(1+\log k)\ a_i(\mathcal{H}_i(\delta_i u_i), \mathcal{H}_i(\delta_i u_i)). \tag{9.2}$$

We next consider a substructure Ω_j that only has an edge $E = E^i = E^j$ in common with Ω_i, with vertices $V^{j1} = V^{i1}$ and $V^{j2} = V^{i2}$. We note that, according to the decompositions (7.1) and (7.2), u has only a wirebasket component on Ω_j,

$$u = u_{Wj} = u_{Vj1} + u_{Vj2} + u_{Ej},$$

which vanishes at all nodes in $\Gamma_{j,h}$ except at those on the closure $\overline{E^j}$. We then have

$$a_j(u,u) \leqslant 3a_j(u_{Vj1}, u_{Vj1}) + 3a_j(u_{Vj2}, u_{Vj2}) + 3a_j(u_{Ej}, u_{Ej}).$$

For the two vertex components, we can proceed as before and find similar bounds to (9.2). For the edge component, we use Lemma 7.1, the definition of $\|\cdot\|_{h,E^j}$ and the fact that δ_i

is constant at all the nodal points in E_h. We find

$$a_j(u_{Ej}, u_{Ej}) = \rho_j |u_{Ej}|^2_{1,\Omega_j} \leqslant C \frac{\rho_j}{(1-\sigma)^2} \|u_{Ej}\|^2_{h,Ej} \leqslant C \frac{\rho_j \, \delta_{i,E}^{-2}}{(1-\sigma)^2} \|\delta_i u_{Ej}\|^2_{h,Ej},$$

where $\delta_{i,E}$ is the constant value of δ_i on E. Thanks to Lemma 8.3, the norm $\|\cdot\|_{h,Ej}$ associated with Ω_j can be bounded by $\|\cdot\|_{h,Ei}$. In addition, we can apply Lemmas 7.5 and 7.3 and find

$$\begin{aligned}
\rho_i \|\delta_i u_{Ej}\|^2_{h,Ej} &\leqslant C(1-\sigma)^{-2} \rho_i \|(\delta_i u_i)_{Ei}\|^2_{h,Ei} \leqslant C(1-\sigma)^{-2} \rho_i \|\mathcal{H}_i(\delta_i u_i)\|^2_{h,Ei} \\
&\leqslant C(1-\sigma)^{-4}(1+\log k)\ \rho_i \|\mathcal{H}_i(\delta_i u_i)\|^2_{1,\Omega_i} \\
&= C(1-\sigma)^{-4}(1+\log k)\ (a_i(\mathcal{H}_i(\delta_i u_i), \mathcal{H}_i(\delta_i u_i)) + \rho_i H_i^{-2} \|\mathcal{H}_i(\delta_i u_i)\|^2_{0,\Omega_i}).
\end{aligned}$$

As before, the L^2 component in the last term can be bounded by the local bilinear form $a_i(\cdot,\cdot)$, thanks to a Poincaré or a Friedrichs inequality. Combining these two estimates and using (9.1), we find

$$a_j(u_{Ej}, u_{Ej}) \leqslant C(1-\sigma)^{-6}(1+\log k)\ a_i(\mathcal{H}_i(\delta_i u_i), \mathcal{H}_i(\delta_i u_i)). \tag{9.3}$$

We next consider a substructure Ω_j that shares a face F and thus also the edges and vertices that lie on ∂F. We note that on Ω_j, u can be decomposed as

$$u = u_{Wj} + u_F.$$

We have

$$a_j(u,u) = \rho_j |u|^2_{1,\Omega_j} \leqslant 2\rho_j (|u_{Wj}|^2_{1,\Omega_j} + |u_F|^2_{1,\Omega_j}).$$

The wirebasket component can be bounded as before; cf. (9.2) and (9.3). For the face component we first note that the function δ_i is equal to a constant value $\delta_{i,F}$ at all nodal points inside F. Using (9.1), we can then write

$$\rho_j |u_F|^2_{1,\Omega_j} = \rho_j \delta_{i,F}^{-2} \, |\mathcal{H}_j(\delta_i u_F)|^2_{1,\Omega_j} \leqslant \rho_i |\mathcal{H}_j(\delta_i u_F)|^2_{1,\Omega_j}.$$

Using Corollary 8.2 and Lemma 7.7 yields

$$\begin{aligned}
|\mathcal{H}_j(\delta_i u_F)|^2_{1,\Omega_j} &\leqslant C(1-\sigma)^{-1} |\mathcal{H}_i(\delta_i u_F)|^2_{1,\Omega_i} \\
&\leqslant C(1-\sigma)^{-5} \left(1 + \log\left(\frac{k}{1-\sigma}\right)\right)^2 ||u||^2_{1,\Omega_i}.
\end{aligned}$$

Combining the last two estimates and using a Poincaré or a Friedrichs inequality, we find

$$a_j(u_F, u_F) \leqslant C\,(1-\sigma)^{-5} \left(1 + \log\left(\frac{k}{1-\sigma}\right)\right)^2 a_i(\mathcal{H}_i(\delta_i u), \mathcal{H}_i(\delta_i u)). \tag{9.4}$$

We finally need to consider the energy of u in Ω_i, $a_i(u,u)$. We note that we can decompose u on Ω_i according to (7.1). The wirebasket and the face components can be bounded as before. Summing over i and the neighbouring subdomains, we then find

$$a(u,u) \leqslant \frac{C}{(1-\sigma)^6} \left(1 + \log\left(\frac{k}{1-\sigma}\right)\right)^2 \left(\sum_{V^{ij}} 1 + \sum_{E^{ij}} 1 + \sum_{F^{ij}} 1\right) a_i(\mathcal{H}_i(\delta_i u), \mathcal{H}_i(\delta_i u)).$$

TABLE 1 *Balancing Neumann–Neumann algorithm*

1. Initialize

$$u_0 = R_0^T S_0^{-1} R_0 g_\Gamma + \tilde{w}, \quad \tilde{w} \in \text{Range}(I - P_0)$$
$$q_0 = g_\Gamma - S u_0$$

2. Iterate $j = 1, 2, \ldots$ until convergence

Project: $w_{j-1} = (I - P_0^T) q_{j-1}$

Precondition: $z_{j-1} = \sum_{i=1}^{N} R_i^T D_i S_i^\dagger D_i R_i \, w_{j-1}$

Project: $y_{j-1} = (I - P_0) z_{j-1}$

$\beta_j = \langle y_{j-1}, w_{j-1} \rangle / \langle y_{j-2}, w_{j-2} \rangle \quad [\beta_1 = 0]$

$p_j = y_{j-1} + \beta_j p_{j-1} \quad [p_1 = y_0]$

$\alpha_j = \langle y_{j-1}, w_{j-1} \rangle / \langle p_j, S p_j \rangle$

$u_j = u_{j-1} + \alpha_j p_j$

$q_j = q_{j-1} - \alpha_j S p_j$

Since the partition $\mathcal{T}_m$ is shape-regular, the number of subdomains to which an edge or a vertex may belong is bounded. We finally obtain

$$\omega \leqslant C (1 - \sigma)^{-6} \left(1 + \log\left(\frac{k}{1 - \sigma}\right)\right)^2.$$

Since in practice σ is bounded away from one, we obtain the same bound as for Neumann–Neumann methods for p finite-element approximations on shape-regular meshes

$$\kappa(P_{NN}) \leqslant C (1 + \log k)^2;$$

(see e.g. Pavarino, 1997). We stress the fact that the constants in the last two estimates are independent of the coefficients ρ_i and the refinement level n (and thus of the aspect ratio of the mesh $\mathcal{T}_{bl}^{n,\sigma}$).

10. Numerical results

The purpose of this section is to present two numerical experiments in order to validate our analysis on some medium-size problems. A more detailed and thorough study will be presented in Toselli & Vasseur (2003b).

The balancing Neumann–Neumann method of Section 6 can be implemented as a projected preconditioned conjugate gradient algorithm and is shown in Table 1 (see Toselli & Vasseur, 2003c for more details). In this table $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product.

It is easy to show that $w_j = q_j$ thanks to the choice of the initial guess, and the first projection step can therefore be omitted. In addition, the application of the pseudoinverses

$S_i^\dagger$ can be carried out by applying the pseudoinverses of the original matrices $A^{(i)}$, cf. (6.1), which amounts to solving local Neumann problems on the substructures (see Smith *et al.*, 1996, Section 4.2.1 for details). The total amount of work for each step consists of the solution of one coarse problem (application of S_0^{-1}), one Neumann problem (application of $S_i^\dagger$) and two Dirichlet problems (application of S for P_0 and for the calculation of the new search direction) on each subdomain. The most expensive parts of the methods are the factorizations of the local matrices $A^{(i)}$ and $A_{II}^{(i)}$, and of the global S_0. The matrices $A^{(i)}$ and $A_{II}^{(i)}$ have roughly the same size.

We remark that the amount of work per step of the unpreconditioned conjugate gradient algorithm for the Schur complement system (6.2) amounts to solving one Dirichlet problem on each substructure (one application of S for the calculation of the new search direction). The rate of convergence however deteriorates very fast with the problem size. A more detailed numerical study on the performance and cost of our algorithm will be performed in Toselli & Vasseur (2003b).

Our first numerical experiment targets the efficiency of the Neumann–Neumann preconditioner for a Laplace problem defined on a boundary layer mesh (corner refinement), whereas the second one is a standard domain decomposition test case defined on a uniform mesh. In both experiments, the conjugate gradient iteration is stopped after a reduction of the Euclidean norm of the initial residual of 10^{-14} and homogeneous Dirichlet boundary conditions have been used.

10.1 *Laplace problem on a boundary layer mesh*

We consider approximations on the unit cube $\Omega = (0, 1)^3$. We choose $\rho \equiv 1$ and the right-hand side $f \equiv 1$. The macromesh $\mathcal{T}_m$ consists of $N \times N \times N$ cubic substructures. Geometric refinement is performed towards the three edges $x = 0$, $y = 0$, and $z = 0$, with $\sigma = 0{\cdot}5$; see Fig. 5, left. Given a polynomial degree k, we choose $n = k$ as is required for robust exponential convergence (see e.g. Andersson *et al.*, 1995; Babuška & Guo, 1996).

We note that even for moderate values of k and N, extremely large linear systems are obtained; cf. Tables 2 and 3. Huge local blocks need to be inverted, both for the application of S (solution of local Dirichlet problems) and the preconditioner (solution of local Neumann problems). Due to memory limitations in our Matlab implementation, direct solvers could not always be employed and thus we have employed approximate solvers for local Dirichlet and Neumann problems. We refer to Smith *et al.* (1996, Section 4.4) for details on the implementation. In particular, we have used a conjugate gradient iteration with an incomplete Cholesky factorization with drop tolerance 10^{-3} for all local problems. The iteration is stopped after a reduction of the initial residual of a factor 10^{-3} or after 20 iteration steps. In the sequel, we denote by NN (inexact) the resulting balancing Neumann–Neumann method with this strategy for the approximate solvers. An exact variant denoted by NN (exact) is derived, when solving all the local subproblems now up to machine precision with the same iterative solver as in the inexact case. Our numerical results show that the theoretical bounds for the case of exact solvers in Lemma 6.2 remain valid in this case; cf. Tables 2 and 3.

For a fixed partition into substructures with $N = 3$, Table 2 shows the size of the original problem, the iteration count, the estimated maximum and minimum eigenvalues,

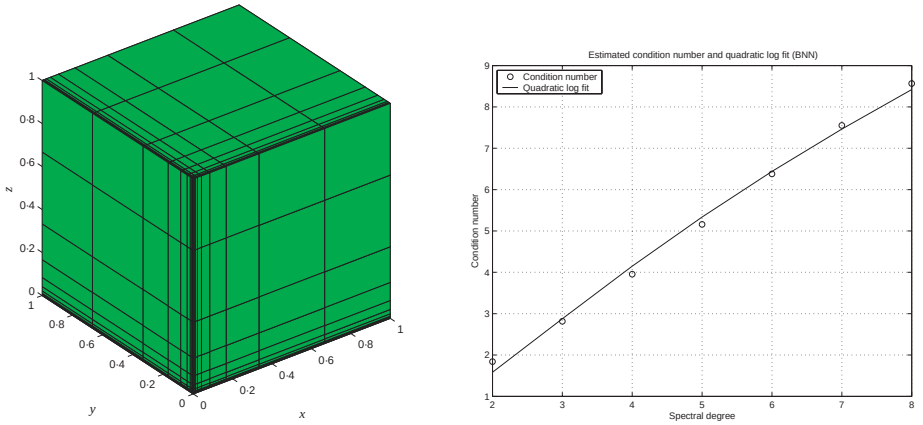


FIG. 5. Geometric refinement towards one corner ($N = 3$, $\sigma = 0{\cdot}5$, and $n = 6$), left, and estimated condition numbers (circles) from Table 2 (inexact variant) and least-square second-order logarithmic polynomial fit (solid line) versus k, right.

TABLE 2 *Conjugate gradient method for the global system with Neumann–Neumann preconditioner with inexact and exact solvers: iteration counts, maximum and minimum eigenvalues, and condition numbers, versus the polynomial degree, for the case of a fixed partition. The size of the original problem is also reported. Fixed number of subdomains* ($N = 3$)

		NN (inexact)				NN (exact)			
k	Size	It	λ_{max}	λ_{min}	κ	It	λ_{max}	λ_{min}	κ
2	1331	15	1·8379	1	1·8379	13	1·6255	1·00002	1·6255
3	6859	20	2·8165	0·99997	2·8166	18	2·8165	1·00001	2·8161
4	24389	25	3·9507	0·99947	3·9528	21	3·9506	1·00002	3·9498
5	68921	29	5·1507	0·99799	5·1611	25	5·1507	1·00002	5·1493
6	166375	34	6·3675	0·99801	6·3803	28	6·3675	1·00002	6·3658
7	357911	38	7·5082	0·99395	7·5540	32	7·5067	1·00002	7·5065
8	704969	40	8·5298	0·99574	8·5663	34	8·5064	1·00002	8·5062

and the condition number for different values of k for both inexact and exact variants. We note that the minimum eigenvalue is close to one when using inexact solvers; see Lemma 6.2. In addition, a moderate growth of the maximum eigenvalue is observed with k; such growth is consistent with the quadratic bound in Lemma 6.3; see Fig. 5, right. Using inexact solvers for the local subproblems induces a moderate increase of number of iterations. Nevertheless, quite satisfactory condition numbers are still obtained, see Table 2.

We next consider the same problem, and fix the polynomial degree $k = 4$. Table 3 shows the results for different values of N. In both variants, the iteration counts, and the smallest and largest eigenvalues appear to be bounded independently of the number of subdomains. We note that when the number of subdomains increases, the number of iterations to reach the convergence criterion for both variants is nearly identical.

TABLE 3 *Conjugate gradient method for the global system with Neumann–Neumann preconditioner with inexact and exact solvers: iteration counts, maximum and minimum eigenvalues, and condition numbers, versus the number of substructures, for the case of a fixed polynomial degree and partitions into $N \times N \times N$ substructures. The size of the original problem is also reported. Fixed spectral degree $k = 4$*

		NN (inexact)				**NN (exact)**			
N	Size	It	$\lambda_{\max}$	$\lambda_{\min}$	κ	It	$\lambda_{\max}$	$\lambda_{\min}$	κ
2	15625	18	2·6417	0·99929	2·6436	15	2·6412	1·0003	2·6406
3	24389	25	3·9507	0·99947	3·9528	21	3·9506	1·0002	3·9498
4	35937	28	4·1084	0·99934	4·1111	25	4·1082	1·0002	4·1074
5	50653	29	4·1378	0·99940	4·1402	26	4·1375	1·0002	4·1369
6	68921	30	4·1492	0·99945	4·1515	28	3·5746	1·0002	3·5741
7	91125	30	4·1555	0·99952	4·1575	28	3·6133	1·0001	3·6128
8	117649	30	4·1593	0·99955	4·1612	29	3·6289	1·0001	3·6284
9	148877	30	4·1618	0·99962	4·1634	29	3·6475	1·0001	3·6470
10	185193	30	4·1636	0·99970	4·1648	29	3·6582	1·0001	3·6577

Nevertheless, the difference on the condition number estimates is more pronounced than in Table 2.

10.2 *Laplace problem with jump coefficients*

The theoretical bound for the condition number in Lemma 6.3 is independent of arbitrary jumps on the coefficients between the substructures. The purpose of this numerical experiment is to check this property. In consequence, the coefficient ρ possibly changes between the substructures by orders of magnitudes. The right-hand side is $f \equiv 1$. Given a partition of $\Omega = (0, 1)^3$ into $N \times N \times N$ cubic substructures ($\mathcal{T} = \mathcal{T}_m = N \times N \times N$), a checkerboard distribution on this partition is considered for ρ which is equal to either ρ_1 or ρ_2 as in Mandel & Brezina (1996). Inexact solvers for the Dirichlet and Neumann problems have been considered.

For a fixed partition into substructures with $N = 3$ and for fixed jumps between the substructures with $\rho_1 = 10^{-3}$ and $\rho_2 = 10^3$, we have investigated the behaviour of the condition number of the preconditioned operator versus the polynomial degree k. This behaviour is shown in Fig. 6 and is consistent with the quadratic bound in Lemma 6.3.

For a fixed partition into substructures with $N = 3$ and for a fixed polynomial degree $k = 4$, we have investigated the influence of the jump ρ_2/ρ_1 on the convergence behaviour of the balancing Neumann–Neumann method. ρ_1 is fixed to 1, whereas ρ_2 is varying from 1 to 10^6. A checkerboard distribution has also been used. The results are presented in Table 4. The number of preconditioned CG iterations in order to satisfy the stopping criterion is bounded independently of the ratio ρ_2/ρ_1, in agreement with the bound for the case of exact solvers in Lemma 6.3.

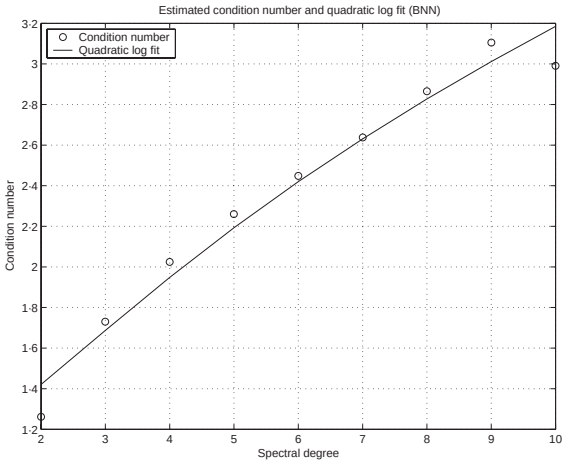


FIG. 6. Laplace problem with jump coefficients. Case of $\rho_1 = 10^{-3}$ and $\rho_2 = 10^3$. Fixed partition $3 \times 3 \times 3$. Estimated condition numbers (circles) and least-square second order logarithmic polynomial (solid line) versus the spectral degree for the balancing Neumann–Neumann method (inexact variant).

TABLE 4 *Laplace problem with jump coefficients. Case of $k = 4$ and $\rho_1 = 1$. Conjugate gradient method for the global system with balancing Neumann–Neumann method (inexact solvers): iteration counts, maximum and minimum eigenvalues, and condition numbers versus ρ_2. Fixed number of subdomains ($N = 3$)*

	NN (inexact)			
ρ_2	It	$\lambda_{\max}$	$\lambda_{\min}$	κ
1	15	2·1153	1	2·1153
10	15	2·1185	0·99999	2·1186
10^2	15	2·0370	1	2·0370
10^3	14	2·0262	1	2·0262
10^4	14	2·0251	0·99991	2·0253
10^5	17	2·0275	0·96406	2·1031
10^6	16	2·0266	0·98234	2·0630

11. Concluding remarks

As for the analysis in Toselli & Vasseur (2003a), some important issues still need to be addressed. We refer to our previous work for a full discussion of these issues.

Our analysis is restricted to approximations that employ nodal basis functions on the Gauss–Lobatto nodes. Indeed, for three-dimensional shape-regular meshes good performance of iterative substructuring methods is in general ensured only if these basis functions are employed and for more general p or hp version finite-element

approximations many important issues remain to be solved even for shape-regular meshes (see e.g. Sherwin & Casarin, 2001 and the references therein).

The Dirichlet and Neumann problems that we need to solve (S_i and $S_i^\dagger$) can be potentially very large. Approximate local solvers can be employed for iterative substructuring methods (see e.g. Smith *et al.*, 1996; Klawonn & Widlund, 2000) and some have been proposed in Korneev *et al.* (2002) for hp-approximations. In our numerical experiments, we have employed a conjugate gradient iteration with an incomplete Choleski preconditioner. However, we believe that the tensor product structure of corner, edge and face patches can be exploited. This is left to a future work.

We believe that the analysis and/or the development of iterative substructuring methods for general meshes with hanging nodes still need to be fully addressed. These meshes are widely used in practice. There is no straightforward way of defining Neumann–Neumann or FETI algorithms when hanging nodes lie on the interface Γ (see Toselli & Vasseur, 2003a, Remark 6.1 for more details).

Finally, our analysis has been carried out for the model problem (2.1), which indeed does not exhibit boundary layers. As for the two-dimensional problems in Toselli & Vasseur (2003a,c), numerical results show that our algorithms are robust when applied to certain singularly perturbed problems. Extensive numerical results will be presented in Toselli & Vasseur (2003b).

Acknowledgements

The authors are grateful to Christoph Schwab and Olof Widlund for enlightening discussions of their work. This work was partially supported by the Swiss National Science Foundation under Project 20-63397.00.

References

AINSWORTH, M. (1996a) A hierarchical domain decomposition preconditioner or $h-p$ finite element approximation on locally efined meshes. *SIAM J. Sci. Comput.*, **17**, 1395–1413.

AINSWORTH, M. (1996b) A preconditioner based on domain decomposition for hp-FE approximation on quasi–uniform meshes. *SIAM J. Numer. Anal.*, **33**, 1358–1376.

AINSWORTH, M. & SHERWIN, S. (1999) Domain decomposition preconditioners for p and hp finite element approximation of Stokes equations. *Comput. Methods Appl. Mech. Engng*, **175**, 243–266.

ANDERSSON, B., FALK, U., BABUŠKA, I. & VON PETERSDORFF, T. (1995) Reliable stress and fracture mechanics analysis of complex aircraft components using a hp-version FEM. *Int. J. Numer. Methods Engng*, **38**, 2135–2163.

BABUŠKA, I. & GUO, B. (1996) Approximation properties of the hp-version of the finite element method. *Comput. Methods Appl. Mech. Engng*, **133**, 319–346.

BERNARDI, C. & MADAY, Y. (1997) Spectral methods. *Handbook of Numerical Analysis*, Vol. V, Part 2. North-Holland: Amsterdam, pp. 209–485.

BHARDWAJ, M., DAY, D., FARHAT, C., LESOINNE, M., PIERSON, K. & RIXEN, D. (2000) Application of the FETI method to ASCI problems: Scalability results on one thousand processors and discussion of highly heterogeneous problems. *Int. J. Numer. Methods Engng*, **47**, 513–535.

BICA, B. (1997) Iterative substructuring algorithms for the p-version finite element method for elliptic problems. *Ph.D. Thesis*, Courant Institute, New York University.

CANUTO, C. (1994) Stablization of spectral methods by finite element bubble functions. *Comput. Methods Appl. Mech. Engng.* (Proc. ICOSAHOM 92, Montpellier, June 1992) 116, pp. 13–26.

CASARIN, M. A. (1996) Schwarz preconditioners for spectral and mortar finite element methods with applications to incompressible fluids. *Ph.D. Thesis* Courant Institute of Mathematical Sciences, March 1996. *Technical Report* 671, Department of Computer Science, New York University, URL: `file://cs.nyu.edu/pub/tech-reports/tr717.ps.gz`

DRYJA, M., SARKIS, M. V. & WIDLUND, O. B. (1996) Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions. *Numer. Math.*, **72**, 313–348.

DRYJA, M. & WIDLUND, O. B. (1995) Schwarz methods of Neumann–Neumann type for three-dimensional elliptic finite element problems. *Commun. Pure Appl. Math.*, **48**, 121–155.

FARHAT, C. & ROUX, F.-X. (1994) Implicit parallel processing in structural mechanics. *Computational Mechanics Advances*, Vol. 2. (J. Oden Tinsley. ed.). North-Holland: pp. 1–124.

GUO, B. & CAO, W. (1997) Additive Schwarz methods for the hp version of the finite element method in two dimensions. *SIAM J. Sci. Comput.*, **18**, 1267–1288.

GUO, B. & CAO, W. (1998) An additive Schwarz method for the hp-version of the finite element method in three dimensions. *SIAM J. Numer. Anal.*, **35**, 632–654.

KLAWONN, A. & WIDLUND, O. B. (2000) A domain decomposition method with Lagrange multipliers and inexact solvers for linear elasticity. *SIAM J. Sci. Comput.*, **22**, 1199–1219.

KLAWONN, A. & WIDLUND, O. B. (2001) FETI and Neumann–Neumann iterative substructuring methods: connections and new results. *Commun. Pure Appl. Math.*, **54**, 57–90.

KORNEEV, V., FLAHERTY, J. E., ODEN, J. T. & FISH, J. (2002) Additive Schwarz algorithms for solving hp-version finite element systems on triangular meshes. *Appl. Numer. Math.*, **43**, 399–421.

LE TALLEC, P. (1994) Domain decomposition methods in computational mechanics. *Computational Mechanics Advances*, Vol. 1. (J. Oden Tinsley. ed.). North-Holland: pp. 121–220.

LE TALLEC, P. & PATRA, A. (1997) Non-overlapping domain decomposition methods for adaptive hp approximations of the Stokes problem with discontinuous pressure fields. *Comput. Methods Appl. Mech. Engng*, **145**, 361–379.

MANDEL, J. (1989) Efficient domain decomposition preconditioning for the p-version finite element method in three dimensions. *Technical Report*, Computational Mathematics Group University of Colorado at Denver.

MANDEL, J. (1990a) Hierarchical preconditioning and partial orthogonalization for the p-version finite element method. *3rd Int. Symp. on Domain Decomposition Methods for Partial Differential Equations* (Houston, Texas, March 20–22, 1989). (T. F. Chan, R. Glowinski, J. Périaux & O. Widlund, eds). Philadelphia, PA: SIAM.

MANDEL, J. (1990b) Two-level domain decomposition preconditioning for the p-version finite element version in three dimensions. *Int. J. Numer. Methods Engng*, **29**, 1095–1108.

MANDEL, J. & BREZINA, M. (1996) Balancing domain decomposition for problems with large jumps in coefficients. *Math. Comput.*, **65**, 1387–1401.

MELENK, J. M. (2002) On condition numbers in hp-FEM with Gauss–Lobatto-based shape functions. *J. Comput. Appl. Math.*, **139**, 21–48.

MELENK, J. M. & SCHWAB, C. (1998) hp–FEM for reaction–diffusion equations. I: robust exponential convergence. *SIAM J. Numer. Anal.*, **35**, 1520–1557.

NEČAS, J. (1967) *Les Méthodes Directes en Théorie des Equations Elliptiques*. Prague: Academia.

ODEN, J. T., PATRA, A. & FENG, Y. (1997) Parallel domain decomposition solver for adaptive hp

finite element methods. *SIAM J. Numer. Anal.*, **34**, 2090–2118.

PAVARINO, L. F. (1994) Additive Schwarz methods for the p-version finite element method. *Numer. Math.*, **66**, 493–515.

PAVARINO, L. F. (1997) Neumann–Neumann algorithms for spectral elements in three dimensions. *RAIRO Math. Modell. Numer. Anal.*, **31**, 471–493.

PAVARINO, L. F. (1998) Preconditioned mixed spectral element ethods for elasticity and Stokes problems. *SIAM J. Sci. Comput.*, **19**, 1941–1957.

PAVARINO, L. F. & WARBURTON, T. (2000) Overlapping Schwarz methods for unstructured spectral elements. *J. Comput. Phys.*, **160**, 298–317.

PAVARINO, L. F. & WIDLUND, O. B. (1996) A polylogarithmic bound for an iterative substructuring method for spectral elements in three dimensions. *SIAM J. Numer. Anal.*, **33**, 1303–1335.

PAVARINO, L. F. & WIDLUND, O. B. (1997) Iterative substructuring methods for spectral elements: Problems in three dimensions based on numerical quadrature. *Comput. Math. Applic.*, **33**, 193–209.

SARKIS, M. V. (September 1994) Schwarz preconditioners for elliptic problems with discontinuous coefficients using conforming and non-conforming elements. *Ph.D. Thesis*, Courant Institute, *Technical Report* 671, Department of Computer Science, New York University, URL: `file://cs.nyu.edu/pub/tech-reports/tr671.ps.Z`

SCHWAB, C. & SURI, M. (1996) The p and hp version of the finite element method for problems with boundary layers. *Math. Comput.*, **65**, 1403–1429.

SCHWAB, C., SURI, M. & XENOPHONTOS, C. A. (1998) The hp-FEM for problems in mechanics with boundary layers. *Comput. Methods Appl. Mech. Engng*, **157**, 311–333.

SHERWIN, S. J. & CASARIN, M. A. (2001) Low energy bases preconditioning for elliptic substructured solvers based on spectral/hp element discretizations. *J. Comput. Phys.*, **171**, 1–24.

SMITH, B. F., BJØRSTAD, P. E. & GROPP, W. D. (1996) *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge: Cambridge University Press.

TOSELLI, A. & VASSEUR, X. (2003a) Neumann–Neumann and FETI preconditioners for hp-approximations on geometrically refined boundary layer meshes in two dimensions. *Technical Report* 02–15, (*Seminar für Angewandte Mathematik, ETH, Zürich, 2002*). Submitted to *Numerische Mathematik*.

TOSELLI, A. & VASSEUR, X. (2003b) A numerical study on Neumann–Neumann and FETI methods for hp-approximations on geometrically refined boundary layer meshes II. *Three-dimensional problems*, in preparation.

TOSELLI, A. & VASSEUR, X. (2003c) A numerical study on Neumann–Neumann and FETI methods for hp-approximations on geometrically refined boundary layer meshes in two dimensions. *Technical Report* 02–20, (*Seminar für Angewandte Mathematik, ETH, Zürich, 2002*). To appear in *Comput. Methods Appl. Mech. Engng.*

B.3. A flexible Generalized Conjugate Residual method with inner orthogonalization and deflated restarting

SIAM J. MATRIX ANAL. APPL.
Vol. 32, No. 4, pp. 1212–1235
© 2011 Society for Industrial and Applied Mathematics

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

A FLEXIBLE GENERALIZED CONJUGATE RESIDUAL METHOD WITH INNER ORTHOGONALIZATION AND DEFLATED RESTARTING*

L. M. CARVALHO†, S. GRATTON‡, R. LAGO§, AND X. VASSEUR¶

Abstract. This work is concerned with the development and study of a minimum residual norm subspace method based on the generalized conjugate residual method with inner orthogonalization (GCRO) method that allows flexible preconditioning and deflated restarting for the solution of non-symmetric or non-Hermitian linear systems. First we recall the main features of flexible generalized minimum residual with deflated restarting (FGMRES-DR), a recently proposed algorithm of the same family but based on the GMRES method. Next we introduce the new inner-outer subspace method named FGCRO-DR. A theoretical comparison of both algorithms is then made in the case of flexible preconditioning. It is proved that FGCRO-DR and FGMRES-DR are algebraically equivalent if a collinearity condition is satisfied. While being nearly as expensive as FGMRES-DR in terms of computational operations per cycle, FGCRO-DR offers the additional advantage to be suitable for the solution of sequences of slowly changing linear systems (where both the matrix and right-hand side can change) through subspace recycling. Numerical experiments on the solution of multidimensional elliptic partial differential equations show the efficiency of FGCRO-DR when solving sequences of linear systems.

Key words. flexible or inner-outer Krylov subspace methods, variable preconditioning, deflation, iterative solver

AMS subject classifications. 65F10, 65N22, 15A06

DOI. 10.1137/100786253

1. Introduction.

In recent years, several authors studied inner-outer Krylov subspace methods that allow variable preconditioning for the iterative solution of large sparse linear systems of equations. One of the first papers describing a subspace method with variable preconditioning is due to Axelsson and Vassilevski, who proposed the generalized conjugate gradient method [1]. See also [2, section 12.3] for additional references. Since then, numerous methods have been proposed to address the symmetric, nonsymmetric, or non-Hermitian cases; these include flexible conjugate gradient [20], flexible GMRES (FGMRES) [24], flexible QMR [31], and GMRESR [34], among others. This class of methods is required when preconditioning with a different (possibly nonlinear) operator at each iteration of a subspace method is considered. This notably occurs when adaptive preconditioners using information obtained from previous iterations [3, 14] are used or when inexact solutions of the preconditioning system using, e.g., adaptive cycling strategy in multigrid [19] or approximate interior solvers in domain decomposition methods [32, section 4.3] are considered. The latter

*Received by the editors February 19, 2010; accepted for publication (in revised form) by Y. Saad August 25, 2011; published electronically November 1, 2011.
http://www.siam.org/journals/simax/32-4/78625.html
†Department of Applied Mathematics, IME-UERJ, R. S. F. Xavier, 524, 629D, 20559-900, Rio de Janeiro, RJ, Brazil (luizmc@gmail.com). This author's work was supported by grant CNPq-473420/2007-4, coordinated by Professor Nelson Maculan.
‡INPT-IRIT, University of Toulouse and ENSEEIHT, 2 rue Camichel, BP 7122, F-31071 Toulouse Cedex 7, France, and CERFACS, 42, Avenue Gaspard Coriolis, F-31057 Toulouse Cedex 1, France (serge.gratton@enseeiht.fr).
§CERFACS, 42, Avenue Gaspard Coriolis, F-31057 Toulouse Cedex 1, France (lago@cerfacs.fr).
¶CERFACS and HiePACS project joint INRIA-CERFACS Laboratory, 42, Avenue Gaspard Coriolis, F-31057 Toulouse Cedex 1, France (vasseur@cerfacs.fr).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

situation is frequent when solving very large systems of linear equations resulting from the discretization of partial differential equations in three dimensions. Thus flexible Krylov subspace methods have gained a considerable interest in the recent years and are the subject of both theoretical and numerical studies [27]. We refer the reader to [29, section 10] for additional comments on flexible methods.

When nonvariable preconditioning is considered, the full GMRES method [23] is often chosen for the solution of nonsymmetric or non-Hermitian linear systems because of its robustness and its minimum residual norm property [26]. Nevertheless to control both the memory requirements and the computational cost of the orthogonalization scheme, restarted GMRES is preferred; it corresponds to a scheme where the maximal dimension of the approximation subspace is fixed. It means in practice that the orthonormal basis built is thrown away at the end of the cycle. Since some information is discarded at the restart, the convergence may stagnate and is expected to be slower compared to full GMRES. Nevertheless to retain the convergence rate a number of techniques have been proposed; they fall in the class of augmented and deflated methods; see, e.g., [4, 10, 11, 16, 25]. Deflated methods compute spectral information at a restart and use this information to improve the convergence of the subspace method. One of the most recent procedures based on a deflation approach is GMRES with deflated restarting (GMRES-DR) [18]. This method reduces to restarted GMRES when no deflation is applied, but may provide a much faster convergence than restarted GMRES for well-chosen deflation spaces as described in [18].

Quite recently a new minimum residual norm subspace method based on GMRES allowing deflated restarting and variable preconditioning has been proposed in [15]. It mainly attempted to combine the numerical features of GMRES-DR and the flexibility property of FGMRES. Numerical experiments in [15] have shown the efficiency of FGMRES with deflated restarting (FGMRES-DR) on both academic and industrial examples. In this paper we study a new minimum residual norm subspace method based on the generalized conjugate method with inner orthogonalization (GCRO) [9] allowing deflated restarting and variable preconditioning. It is named flexible generalized conjugate residual method with inner orthogonalization and deflated restarting (FGCRO-DR) and can be viewed as an extension of GCRO-DR [22] to the case of variable preconditioning. A major advantage of FGCRO-DR over FGMRES-DR is its ability to solve sequences of linear systems (where both the left- and right-hand sides can change) through recycling [22]. In [22] Parks et al. mentioned that GCRO-DR and GMRES-DR were algebraically equivalent, i.e., both methods produce the same iterates in exact arithmetic when solving the same given linear system starting from the same initial guess. When variable preconditioning is considered, it seems therefore natural to ask whether FGCRO-DR and FGMRES-DR could also be algebraically equivalent. We address this question in this paper, and the main theoretical developments that are proposed will help us to answer this question. The main contributions of the paper are then twofold. First we prove that FGCRO-DR and FGMRES-DR can be considered algebraically equivalent if a collinearity condition between two certain vectors is satisfied at each cycle. When considering nonvariable preconditioning, these theoretical developments will also allow us to show the algebraic equivalence between GCRO-DR and GMRES-DR that was stated without proof in [22]. Second we carefully analyze the computational cost of FGCRO-DR and show that the proposed method is nearly as expensive as FGMRES-DR in terms of operations per cycle. Furthermore it is explained how to include subspace recycling into FGCRO-DR, and numerical experiments are reported showing the efficiency of FGCRO-DR.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

This paper is organized as follows. In section 2 we introduce the general background of this study. We briefly recall the main properties of FGMRES-DR and then introduce the FGCRO-DR method both from a mathematical and an algorithmic point of view. Section 3 is mainly devoted to the analysis of both flexible methods. Therein we show that both methods can be algebraically equivalent in the flexible case if a certain collinearity condition is satisfied at each cycle. In section 4 we compare FGCRO-DR and FGMRES-DR in terms of computational operations per cycle and storage and discuss the solution of sequences of linear systems through subspace recycling. Finally we draw some conclusions and perspectives in section 5.

2. Flexible Krylov methods with restarting.

2.1. General setting.

Notation. Throughout this paper we denote by $\|.\|$ the Euclidean norm, by $I_k \in \mathbb{C}^{k\times k}$ the identity matrix of dimension k, and by $0_{i\times j} \in \mathbb{C}^{i\times j}$ the zero rectangular matrix with i rows and j columns. Given $N \in \mathbb{C}^{n\times m}$, $\Pi_{N^\perp} = I_n - N\,N^\dagger$ will represent the orthogonal projector onto range$(N)^\perp$, where the superscript $\dagger$ refers to the Moore–Penrose inverse. Finally, given $Z_m = [z_1, \ldots, z_m] \in \mathbb{C}^{n\times m}$, we will usually decompose Z_m into two submatrices defined as $Z_k = [z_1, \ldots, z_k] \in \mathbb{C}^{n\times k}$ and $Z_{m-k} = [z_{k+1}, \ldots, z_m] \in \mathbb{C}^{n\times(m-k)}$.

Setting. We focus on minimum residual norm based subspace methods that allow flexible preconditioning for the iterative solution of

$$Ax = b, \quad A \in \mathbb{C}^{n\times n}, \quad x, b \in \mathbb{C}^n, \tag{2.1}$$

given an initial vector $x_0 \in \mathbb{C}^n$. In this paper A is supposed to be nonsingular. Flexible methods refer to a class of methods where the preconditioner is allowed to vary at each iteration. We refer the reader to, e.g., [29] for a general introduction on Krylov subspace methods and to [29, section 10] and [26, section 9.4] for a review on flexible methods. The minimum residual norm GMRES method [23] has been extended by Saad [24] to allow variable preconditioning. The resulting algorithm known as FGMRES(m) relies on the Arnoldi relation

$$AZ_m = V_{m+1}\bar{H}_m, \tag{2.2}$$

where $Z_m \in \mathbb{C}^{n\times m}$, $V_{m+1} \in \mathbb{C}^{n\times(m+1)}$ has orthonormal columns and $\bar{H}_m \in \mathbb{C}^{(m+1)\times m}$ is upper Hessenberg. We denote by $\mathcal{M}_j$ the preconditioning operator at iteration j and remark that $\mathcal{M}_j$ may be a nonlinear preconditioning function. We will then denote by $\mathcal{M}_j(v)$ the action of $\mathcal{M}_j$ on a vector v. In (2.2), the columns of V_{m+1} form an orthonormal basis of the subspace spanned by the vectors

$$\{r_0, Az_1, \ldots, Az_m\} \quad \text{with} \quad r_0 = b - Ax_0,$$

whereas $Z_m = [z_1, \ldots, z_m]$ and $V_m = [v_1, \ldots, v_m]$ are related by

$$Z_m = [\mathcal{M}_1(v_1), \ldots, \mathcal{M}_m(v_m)] \quad \text{with} \quad v_1 = \frac{r_0}{\|r_0\|}.$$

At the end of the cycle an approximate solution $x_m \in \mathbb{C}^n$ is then found by minimizing the residual norm $\|r_0 - AZ_m y\|$ over the space $x_0 + \text{range}(Z_m)$. Thus we obtain that

$$x_m = x_0 + Z_m y^*,$$

where y^* is the solution of the following least-squares problem of size $(m+1)\times m$:

$$y^* = \text{argmin}_{y\in\mathbb{C}^m}\|r_0 - AZ_m y\| = \text{argmin}_{y\in\mathbb{C}^m}\|\|r_0\|\, e_1 - \bar{H}_m y\|,$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

where e_1 is the first canonical vector of $\mathbb{C}^{m+1}$. Flexible subspace methods with restarting are based on a procedure where the construction of the subspace is stopped after a certain number of steps (denoted by m in this paper with $m < n$). The method is then restarted mainly to control both the memory requirements and the cost of the orthogonalization scheme. In FGMRES(m) the restarting consists in taking as an initial guess the last iterate of the cycle (x_m).

The main focus of this paper is to present minimum residual norm subspace methods with *deflated* restarting that allow *flexible* preconditioning. Deflated restarting aims at determining an approximation subspace of dimension m as a direct sum of two subspaces of smaller dimension, where one of these subspaces will contain relevant spectral information that will be kept for the next cycle. We refer the reader to, e.g., [25] and [29, section 9] for a review of augmented and deflated methods. Flexible methods with deflated restarting will notably satisfy the following flexible Arnoldi relation:

$$AZ_m = V_{m+1}\bar{H}_m \quad \text{with} \quad V_{m+1}^H V_{m+1} = I_{m+1}, \tag{2.3}$$

where $\bar{H}_m \in \mathbb{C}^{(m+1)\times m}$ is not necessarily of upper Hessenberg form. In this paper we call this relation a flexible Arnoldi-like relation due to its similarity to relation (2.2).

Stagnation and breakdown. We refer the reader to [27, section 6] for general comments and a detailed discussion on the possibility of both breakdown and stagnation in flexible inner-outer Krylov subspace methods. Although important, these issues are not addressed in this paper, and we assume that no breakdown occurs in the inner-outer subspace methods that will be proposed.

2.2. Flexible GMRES with deflated restarting. A number of techniques have been proposed to compute spectral information at a restart and use this information to improve the convergence rate of the Krylov subspace methods; see, e.g., [16, 17, 18, 25]. These techniques have been exclusively developed in the case of a fixed preconditioner. GMRES-DR is one of these methods. It focuses on removing (or deflating) the eigenvalues of smallest magnitude. A full subspace of dimension k, $k < m$ (and not only the approximate solution with minimum residual norm) is now retained at the restart, and the success of this approach has been demonstrated in many academic examples [16]. Approximations of eigenvalues of smallest magnitude are obtained by computing harmonic Ritz pairs of A with respect to a certain subspace [18]. We present here a definition of a harmonic Ritz pair equivalent to the one introduced in [21, 30]; it will be of key importance when defining appropriate deflation strategies.

DEFINITION 2.1 (harmonic Ritz pair). *Consider a subspace $\mathcal{U}$ of $\mathbb{C}^n$. Given $B \in \mathbb{C}^{n\times n}$, $\theta \in \mathbb{C}$, and $y \in \mathcal{U}$, (θ, y) is a harmonic Ritz pair of B with respect to $\mathcal{U}$ if and only if*

$$By - \theta\, y \perp B\,\mathcal{U}$$

or, equivalently, for the canonical scalar product,

$$\forall w \in \mathrm{range}(B\,\mathcal{U}) \quad w^H\,(By - \theta\, y) = 0.$$

We call y a harmonic Ritz vector associated with the harmonic Ritz value θ.

As in the case of fixed preconditioning, deflated restarting may also improve the convergence rate of flexible subspace methods. In [15] a deflated restarting procedure has been proposed for the FGMRES algorithm. The ith cycle of the resulting algorithm, called FGMRES-DR, is now briefly described, and we denote by

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$r_0^{(i-1)} = b - Ax_0^{(i-1)}$, V_{m+1}, $\bar{H}_m$, and Z_m the residual and matrices obtained at the end of the $(i-1)$th cycle.

Based on the Arnoldi-like relation (2.3), the deflation procedure proposed in [15, Proposition 1] relies on the use of k harmonic Ritz vectors $Y_k = V_m P_k$ of $AZ_m V_m^H$ with respect to range(V_m), where $Y_k \in \mathbb{C}^{n\times k}$ and $P_k \in \mathbb{C}^{m\times k}$. In Lemma 2.2 shown in [15, Lemma 3.1], we recall a useful relation satisfied by the harmonic Ritz vectors.

LEMMA 2.2. *In FGMRES-DR, the harmonic Ritz vectors are given by* $Y_k = V_m P_k$ *with corresponding harmonic Ritz values* λ_k. $P_k \in \mathbb{C}^{m\times k}$ *satisfies the following relation:*

$$AZ_m P_k = V_{m+1}\left[\begin{bmatrix} P_k \\ 0_{1\times k}\end{bmatrix}, c - \bar{H}_m y^*\right]\begin{bmatrix} \operatorname{diag}(\lambda_1,\ldots,\lambda_k) \\ \alpha_{1\times k}\end{bmatrix}, \tag{2.4}$$

$$AZ_m P_k = [V_m P_k, r_0^{(i-1)}]\begin{bmatrix} \operatorname{diag}(\lambda_1,\ldots,\lambda_k) \\ \alpha_{1\times k}\end{bmatrix}, \tag{2.5}$$

where $r_0^{(i-1)} = V_{m+1}(c - \bar{H}_m y^*)$ *and* $\alpha_{1\times k} = [\alpha_1,\ldots,\alpha_k] \in \mathbb{C}^{1\times k}$.

Next, the QR factorization of the $(m+1)\times(k+1)$ matrix appearing on the right-hand side of relation (2.4) is performed as

$$\left[\begin{bmatrix} P_k \\ 0_{1\times k}\end{bmatrix}, c - \bar{H}_m y^*\right] = QR, \tag{2.6}$$

where $Q \in \mathbb{C}^{(m+1)\times(k+1)}$ has orthonormal columns and $R \in \mathbb{C}^{(k+1)\times(k+1)}$ is upper triangular, respectively. We write the matrix Q obtained in relation (2.6) as

$$Q = \left[\begin{bmatrix} Q_{m\times k} \\ 0_{1\times k}\end{bmatrix}, \frac{\bar{\rho}}{\|\bar{\rho}\|}\right], \tag{2.7}$$

where $Q_{m\times k} \in \mathbb{C}^{m\times k}$ and $\bar{\rho} \in \mathbb{C}^{m+1}$ is defined as

$$\bar{\rho} = \left(I_{m+1} - \begin{bmatrix} Q_{m\times k} \\ 0_{1\times k}\end{bmatrix}\begin{bmatrix} Q_{m\times k} \\ 0_{1\times k}\end{bmatrix}^H\right)(c - \bar{H}_m y^*). \tag{2.8}$$

PROPOSITION 1. *In FGMRES-DR, the flexible Arnoldi relation*

$$A\ Z_k = V_{k+1}\bar{H}_k, \tag{2.9}$$

$$V_{k+1}^H V_{k+1} = I_{k+1}, \tag{2.10}$$

$$\operatorname{range}\left(\left[Y_k, r_0^{(i-1)}\right]\right) = \operatorname{range}(V_{k+1}) \tag{2.11}$$

holds at the i*th cycle with matrices* $Z_k, V_k \in \mathbb{C}^{n\times k}$ *and* $\bar{H}_k \in \mathbb{C}^{(k+1)\times k}$ *defined as*

$$Z_k = Z_m Q_{m\times k}, \tag{2.12}$$

$$V_{k+1} = V_{m+1} Q, \tag{2.13}$$

$$\bar{H}_k = Q^H \bar{H}_m Q_{m\times k}, \tag{2.14}$$

where V_{m+1}, Z_m, *and* $\bar{H}_m$ *refer to matrices obtained at the end of the* $(i-1)$*th cycle.*

Proof. Relations (2.9), (2.10), (2.12), (2.13), and (2.14) have been shown in [15, Proposition 2]. From relations (2.13) and (2.6), respectively, we deduce

$$V_{k+1}R = V_{m+1}\left[\begin{bmatrix} P_k \\ 0_{1\times k}\end{bmatrix}, c - \bar{H}_m y^*\right],$$

$$V_{k+1}R = \left[V_m P_k, r_0^{(i-1)}\right], \tag{2.15}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

which finally shows that range($[Y_k, r_0^{(i-1)}]$) = range(V_{k+1}) since R is supposed to be nonsingular. ∎

FGMRES-DR then carries out $m-k$ Arnoldi steps with flexible preconditioning and starting vector v_{k+1} while maintaining orthogonality to V_k, leading to

$$A\,[z_{k+1},\ldots,z_m] = [v_{k+1},\ldots,v_{m+1}]\,\bar{H}_{m-k} \quad \text{and} \quad V_{m+1}^H\,V_{m+1} = I_{m+1}.$$

We note that $\bar{H}_{m-k} \in \mathbb{C}^{(m-k+1)\times(m-k)}$ is upper Hessenberg. At the end of the ith cycle this gives the flexible Arnoldi-like relation

$$A\,[Z_k, Z_{m-k}] = [V_{m+1}]\,\left[\begin{bmatrix}\bar{H}_k\\ 0_{m-k\times k}\end{bmatrix}\,\begin{bmatrix}B_{k\times m-k}\\ \bar{H}_{m-k}\end{bmatrix}\right],$$

where $V_{m+1} \in \mathbb{C}^{n\times(m+1)}$, $\bar{H}_m \in \mathbb{C}^{(m+1)\times m}$, and $B_{k\times m-k} \in \mathbb{C}^{k\times(m-k)}$ results from the orthogonalization of $[v_{k+2},\ldots,v_{m+1}]$ against V_{k+1}. We note that $\bar{H}_m$ is no longer upper Hessenberg due to the leading dense $(k+1)\times k$ submatrix $\bar{H}_k$. At the end of the ith cycle, an approximate solution $x_0^{(i)} \in \mathbb{C}^n$ is then found by minimizing the residual norm $\|b - A(x_0^{(i-1)} + Z_m y)\|$ over the space $x_0^{(i-1)} + \text{range}(Z_m)$, the corresponding residual being $r_0^{(i)} = b - Ax_0^{(i)}$, with $r_0^{(i)} \in \text{range}(V_{m+1})$. We refer the reader to [15] for the complete derivation of the method and numerical experiments showing the efficiency of FGMRES-DR in both academic and industrial examples.

2.3. Flexible GCRO with deflated restarting. GCRO-DR [22]—a combination of GMRES-DR and GCRO—is a Krylov subspace method that allows deflated restarting and subspace recycling simultaneously. This latter feature is particularly interesting when solving sequences of linear systems with possibly different left- or right-hand sides. As pointed out in [22], GCRO-DR is attractive because any subspace may be recycled. In this paper we restrict the presentation to the case of a single linear system as proposed in (2.1).

GCRO and GCRO-DR belong to the family of inner-outer methods [2, Chap. 12] where the outer iteration is based on GCR, a minimum residual norm method proposed by Eisenstat, Elman, and Schultz [13]. To this end GCR maintains a correction subspace spanned by range(Z_m) and an approximation subspace spanned by range(V_m), where $Z_m, V_m \in \mathbb{C}^{n\times m}$ satisfy

$$\begin{aligned} A\,Z_m &= V_m,\\ V_m^H\,V_m &= I_m. \end{aligned}$$

The optimal solution of the minimization problem $\min\|b - Ax\|$ over the subspace $x_0 + \text{range}(Z_m)$ is then found as $x_m = x_0 + Z_m\,V_m^H\,r_0$. Consequently $r_m = b - A\,x_m$ satisfies

$$r_m = r_0 - V_m\,V_m^H r_0 = \Pi_{V_m^\perp} r_0, \quad r_m \perp \text{range}(V_m).$$

In [9] de Sturler proposed an improvement to GMRESR [34], an inner-outer method based on GCR in the outer part and GMRES in the inner part, respectively. He suggested that the inner iteration takes place in a subspace orthogonal to the outer Krylov subspace. In this inner iteration the projected residual equation

$$\left(I_n - V_m\,V_m^H\right) Az = r_m$$

is solved only approximately. If a minimum residual norm subspace method is used in the inner iteration to solve this projected residual linear system, the residuals over

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

both the inner and outer subspaces are minimized. This leads to the GCRO Krylov subspace method [9]. Numerical experiments [9] indicate that the resulting method may perform better than other inner-outer methods (without orthogonalization) in some cases.

The GCRO method with deflated restarting (named GCRO-DR) based on harmonic Ritz value information was proposed in [22]. An approximate invariant subspace is used for deflation following closely the GMRES-DR method. We refer the reader to [22] for a description of this method, algorithms, and implementation details. We present now a new variant of GCRO-DR that allows flexible preconditioning by explaining the different steps occurring during the ith cycle. Again we denote by $r_0^{(i-1)} = b - Ax_0^{(i-1)}$, V_{m+1}, $\bar{H}_m$, and Z_m the residual and matrices obtained at the end of the $(i-1)$th cycle.

We suppose that a flexible Arnoldi-like relation of type (2.3) holds. As in section 2.2 an important point is to specify which harmonic Ritz information is selected. Given a certain matrix $W_m \in \mathbb{C}^{n\times m}$ to be specified later on, such as range(W_m) = range(V_m), the deflation procedure relies on the use of k harmonic Ritz vectors $Y_k = W_m P_k$ of $AZ_m W_m^\dagger$ with respect to range(W_m), where $Y_k \in \mathbb{C}^{n\times k}$ and $P_k \in \mathbb{C}^{m\times k}$. W_m will notably satisfy a property detailed in Lemma 3.3 and we point out that the calculation of $W_m^\dagger$ is not needed in the practical implementation of the algorithm (see section 4.1.1). In Lemma 2.3 we detail a useful relation satisfied by the harmonic Ritz vectors.

LEMMA 2.3. *In flexible GCRO with deflated restarting (FGCRO-DR), the harmonic Ritz vectors are given by $Y_k = W_m P_k$ with corresponding harmonic Ritz values θ_k. The matrix $P_k = [p_1, \ldots, p_k] \in \mathbb{C}^{m\times k}$ satisfies the following relation:*

$$AZ_m P_k = \left[W_m P_k, r_0^{(i-1)}\right] \begin{bmatrix} \operatorname{diag}(\theta_1, \ldots, \theta_k) \\ \beta_{1\times k} \end{bmatrix}, \tag{2.16}$$

where $r_0^{(i-1)} = V_{m+1}(c - \bar{H}_m y^*)$ and $\beta_{1\times k} = [\beta_1, \ldots, \beta_k] \in \mathbb{C}^{1\times k}$.

Proof. According to Definition 2.1, the harmonic residual vectors $AZ_m W_m^\dagger W_m p_j - \theta_j W_m p_j$ and the residual vector $r_0^{(i-1)} = V_{m+1}(c - \bar{H}_m y^*)$ all belong to a subspace of dimension $m+1$ (spanned by the columns of V_{m+1}) and are orthogonal to the same subspace of dimension m (spanned by the columns of AZ_m subspace of range(V_{m+1})), so they must be collinear. Consequently there exist k coefficients noted $\beta_j \in \mathbb{C}$ with $1 \le j \le k$ such that

$$\forall j \in \{1, \ldots, k\} \quad AZ_m p_j - \theta_j W_m p_j = \beta_j r_0^{(i-1)}. \tag{2.17}$$

Setting $\beta_{1\times k} = [\beta_1, \ldots, \beta_k] \in \mathbb{C}^{1\times k}$, the collinearity expression (2.17) can be written in matrix form as

$$AZ_m P_k = \left[W_m P_k, r_0^{(i-1)}\right] \begin{bmatrix} \operatorname{diag}(\theta_1, \ldots, \theta_k) \\ \beta_{1\times k} \end{bmatrix}.$$

Due to the flexible Arnoldi-like relation (2.3), relation (2.16) can be also expressed as

$$V_{m+1}\bar{H}_m P_k = \left[W_m P_k, r_0^{(i-1)}\right] \begin{bmatrix} \operatorname{diag}(\theta_1, \ldots, \theta_k) \\ \beta_{1\times k} \end{bmatrix}. \tag{2.18}$$

If required, $\beta_{1\times k}$ can be deduced from (2.18) by

$$(c - \bar{H}_m y^*)^H (\bar{H}_m P_k - V_{m+1}^H W_m P_k \operatorname{diag}(\theta_1, \ldots, \theta_k)) = (c - \bar{H}_m y^*)^H (c - \bar{H}_m y^*)\beta_{1\times k}. \tag{2.19}$$

∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Next, the QR factorization of the $(m+1)\times k$ matrix $\bar{H}_m P_k$ appearing in relation (2.18) is performed as $\bar{H}_m P_k = QR$ with $Q \in \mathbb{C}^{(m+1)\times k}$ and $R \in \mathbb{C}^{k\times k}$.

PROPOSITION 2. *In FGCRO-DR, the relation $AZ_k = V_k$ with $V_k^H V_k = I_k$ holds at the ith cycle with matrices $Z_k, V_k \in \mathbb{C}^{n\times k}$ defined as*

$$\begin{aligned} Z_k &= Z_m P_k R^{-1}, \\ V_k &= V_{m+1} Q, \end{aligned}$$

where V_{m+1} and Z_m refer to matrices obtained at the end of the $(i-1)$th cycle. In addition $V_k^H r_0^{(i-1)} = 0$ holds during the ith cycle.

Proof. By using information related to the QR factorization of $\bar{H}_m P_k$ and the flexible Arnoldi relation (2.3) exclusively, we obtain

$$\begin{aligned} A\ Z_k &= AZ_m P_k R^{-1} \\ &= V_{m+1} \bar{H}_m P_k R^{-1} \\ &= V_{m+1} Q \\ &= V_k. \end{aligned}$$

Since both V_{m+1} and Q have orthonormal columns, V_k satisfies $V_k^H V_k = I_k$. Finally since $r_0^{(i-1)}$ is the optimum residual at the $(i-1)$th cycle, i.e., $(AZ_m)^H r_0^{(i-1)} = 0$, we obtain

$$\begin{aligned} P_k^H (AZ_m)^H r_0^{(i-1)} &= 0, \\ (V_{m+1}\bar{H}_m P_k)^H r_0^{(i-1)} &= 0, \\ R^H V_k^H r_0^{(i-1)} &= 0. \end{aligned}$$

This finally shows that $V_k^H r_0^{(i-1)} = 0$ since R is supposed to be nonsingular. ∎

To complement the subspaces, the inner iteration is based on the approximate solution of

$$\left(I_n - V_k\ V_k^H\right) Az = \left(I_n - V_k\ V_k^H\right) r_0^{(i-1)} = r_0^{(i-1)},$$

where the last equality is due to Proposition 2. For that purpose FGCRO-DR then carries out $m-k$ steps of the Arnoldi method with flexible preconditioning, leading to

$$\begin{aligned} \left(I_n - V_k V_k^H\right)\ A\ [z_{k+1}, \ldots, z_m] &= [v_{k+1}, \ldots, v_{m+1}]\ \bar{H}_{m-k}, \\ \left(I_n - V_k V_k^H\right)\ A\ Z_{m-k} &= V_{m-k+1}\ \bar{H}_{m-k} \end{aligned}$$

with $v_{k+1} = r_0^{(i-1)}/||r_0^{(i-1)}||$. At the end of the cycle this gives the flexible Arnoldi-like relation

$$\begin{aligned} A\ [Z_k, Z_{m-k}] &= [V_k, V_{m-k+1}] \begin{bmatrix} I_k & V_k^H\ A\ Z_{m-k} \\ 0_{m-k+1\times k} & \bar{H}_{m-k} \end{bmatrix}, \\ A\ Z_m &= V_{m+1}\ \bar{H}_m, \end{aligned}$$

where $Z_m \in \mathbb{C}^{n\times m}$, $V_{m+1} \in \mathbb{C}^{n\times(m+1)}$, and $\bar{H}_m \in \mathbb{C}^{(m+1)\times m}$. At the end of the ith cycle, an approximate solution $x_0^{(i)} \in \mathbb{C}^n$ is then found by minimizing the residual norm $\|b - A(x_0^{(i-1)} + Z_m y)\|$ over the space $x_0^{(i-1)} + \text{range}(Z_m)$, the corresponding residual being $r_0^{(i)} = b - Ax_0^{(i)}$, with $r_0^{(i)} \in \text{range}(V_{m+1})$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

2.4. Algorithms. Details of the FGCRO-DR method are given in Algorithm 1, where MATLAB-like notations are adopted (for instance, in step 7b, $Q(1:m, 1:k)$ denotes the submatrix made of the first m rows and first k columns of matrix Q noted $Q_{m\times k}$ in (2.7)). For the sake of completeness the FGMRES-DR algorithm has also been described with notation chosen as closely as possible to FGCRO-DR to make code comparison easier. Concerning Algorithm 1 we make the following comments:

- As will be discussed later, the computation of $W_m^\dagger$ in step 5a is not required thanks to the definition of the harmonic Ritz pair (see Definition 2.1).
- As pointed out by Morgan [18] and Parks et al. [22] we might have to adjust k during the algorithm to include both the real and imaginary parts of complex eigenvectors.
- In steps 10a and 10b $\mathcal{M}_j^{(i)}$ denotes the possibly nonlinear preconditioning operator at iteration j during the ith cycle.

Algorithm 1. FGCRO-DR(m,k) and FGMRES-DR(m,k).

1: Choose m, k, tol, and x_0
2: $r_0 = b - Ax_0$, $\beta = \|r_0\|$, $v_1 = r_0/\beta$, $c = \beta e_1$, $i \leftarrow 0$
3: Apply FGMRES(m) to obtain $\bar{H}_m$, Z_m, V_{m+1} such that $AZ_m = V_{m+1}\bar{H}_m$, $y^* = \arg\min_{y\in\mathbb{C}^m} \|c - \bar{H}_m y\|$, $x_0^{(0)} = x_0 + Z_m y^*$, $r_0^{(0)} = b - Ax_0^{(0)} = V_{m+1}(c - \bar{H}_m y^*)$, $W_m = V_m$
4: **while** $\|r_0^{(i)}\| > \|b\| \times tol$ **do** $i \leftarrow i + 1$

FGCRO-DR

5a: Compute k harmonic Ritz vectors of $AZ_m W_m^\dagger$ with respect to range(W_m) and store them in Y_k. Define P_k such that $Y_k = W_m P_k$
6a: $Q\ R = \bar{H}_m P_k$
7a: $W_k = W_m P_k R^{-1}$
8a: $V_k = V_{m+1} Q$
9a: $Z_k = Z_m P_k R^{-1}$
10a: Apply $m-k$ flexible preconditioned Arnoldi steps with $(I_n - V_k V_k^H)A$ and $v_{k+1} = r_0^{(i-1)}/\|r_0^{(i-1)}\|$ such that $(I_n - V_k V_k^H)A\left[z_{k+1},\ldots,z_m\right] = \left[v_{k+1},\ldots,v_{m+1}\right]\bar{H}_{m-k}$ with $z_j = \mathcal{M}_j^{(i)}(v_j)$
11a: Set $\bar{H}_m = \begin{bmatrix} I_k & V_k^H A Z_{m-k} \\ 0_{m-k+1\times k} & \bar{H}_{m-k} \end{bmatrix}$ yielding $A\left[z_1,\ldots,z_m\right] = \left[v_1,\ldots,v_{m+1}\right]\bar{H}_m$ and define $W_m = \left[W_k \quad V_m(1:n, k+1:m)\right]$

FGMRES-DR

5b: Compute k harmonic Ritz vectors of $AZ_m V_m^H$ with respect to range(V_m) and store them in Y_k. Define P_k such that $Y_k = V_m P_k$
6b: $QR = \left[\begin{bmatrix} P_k \\ 0_{1\times k} \end{bmatrix} \quad c - \bar{H}_m y^*\right]$
7b: $\bar{H}_k = Q^H \bar{H}_m Q(1:m, 1:k)$
8b: $V_{k+1} = V_{m+1} Q$
9b: $Z_k = Z_m Q(1:m, 1:k)$
10b: Apply $m-k$ flexible preconditioned Arnoldi steps with A and v_{k+1} while maintaining orthogonality to V_k such that $A\left[z_{k+1},\ldots,z_m\right] = \left[v_{k+1},\ldots,v_{m+1}\right]\bar{H}_{m-k}$ with $z_j = \mathcal{M}_j^{(i)}(v_j)$ and $V_{m+1}^H V_{m+1} = I_{m+1}$
11b: Set $\bar{H}_m = \left[\begin{bmatrix} \bar{H}_k \\ 0_{m-k\times k} \end{bmatrix} \begin{bmatrix} B_{k\times m-k} \\ \bar{H}_{m-k} \end{bmatrix}\right]$ yielding $A\left[z_1,\ldots,z_m\right] = \left[v_1,\ldots,v_{m+1}\right]\bar{H}_m$

12: $y^* = \arg\min_{y\in\mathbb{C}^m} \|c - \bar{H}_m y\|$ with $c = V_{m+1}^H r_0^{(i-1)}$
13: $x_0^{(i)} = x_0^{(i-1)} + Z_m y^*$
14: $r_0^{(i)} = b - Ax_0^{(i)}$
15: **end while**

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

3. Analysis of FGMRES-DR and FGCRO-DR. We compare now the flexible variants of GMRES-DR and GCRO-DR introduced in sections 2.2 and 2.3, respectively. In the following we use $\widetilde{\ }$ to denote quantities related to the FGMRES-DR algorithm, e.g., $\widetilde{Y}_k$ denotes the set of harmonic Ritz vectors computed in the FGMRES-DR algorithm. When analyzing both algorithms we will suppose that identical preconditioning operators are used in steps 10a and 10b, respectively, i.e.,

$$\forall i, \forall j \in \{k+1, \ldots, m\}, \quad \mathcal{M}_j^{(i)}(.) = \widetilde{\mathcal{M}}_j^{(i)}(.). \tag{3.1}$$

3.1. Equivalent preconditioning matrix.

DEFINITION 3.1 (equivalent preconditioning matrix). *Suppose that* $V_p = [v_1, \ldots, v_p] \in \mathbb{C}^{n\times p}$ *and* $Z_p = [\mathcal{M}_1(v_1), \ldots, \mathcal{M}_p(v_p)] \in \mathbb{C}^{n\times p}$ *obtained during a cycle of a flexible method with (standard or deflated) restarting (with* $1 \le p \le m < n$*) are both of full rank, i.e.,* $\operatorname{rank} V_p = \operatorname{rank} Z_p = p$*. We will then denote by* $M_{V_p} \in \mathbb{C}^{n\times n}$ *a nonsingular equivalent preconditioning matrix defined as*

$$Z_p \overset{\text{def}}{=} M_{V_p}\, V_p. \tag{3.2}$$

Such a matrix represents the action of the nonlinear operators $\mathcal{M}_j$ *on the set of vectors* v_j *(with* $j = 1, \ldots, p$*). It can be chosen, e.g., as* $M_{V_p} = [Z_p\ \underline{Z_p}][V_p\ \underline{V_p}]^{-1}$*, where* $\underline{Z_p}$ *(respectively,* $\underline{V_p}$*) denotes an orthogonal complement of* Z_p *(respectively,* V_p*) in* $\mathbb{C}^n$*.*

3.2. Relations between Z_m and W_m and $\widetilde{Z}_m$ and $\widetilde{V}_m$. We denote by $M_{W_m}^{(0)}$ and $\widetilde{M}_{\widetilde{V}_m}^{(0)}$ the equivalent preconditioning matrices used in the initialization phase of both algorithms (step 3 in Algorithm 1). With this notation we remark that the following relations hold:

$$Z_m = M_{W_m}^{(0)} W_m = \widetilde{Z}_m = \widetilde{M}_{\widetilde{V}_m}^{(0)} \widetilde{V}_m. \tag{3.3}$$

We first analyze the relation between $\widetilde{Z}_m$ and $\widetilde{V}_m$.

LEMMA 3.2. *At the end of the* i*th cycle of the FGMRES-DR method,* $\widetilde{Z}_m$ *and* $\widetilde{V}_m$ *satisfy*

$$\widetilde{Z}_m = \widetilde{M}_{\widetilde{V}_m}^{(i)} \widetilde{V}_m = \left[\widetilde{M}_{\widetilde{V}_m}^{(i-1)} \widetilde{V}_k,\ \widetilde{M}_{\widetilde{V}_{m-k}}^{(i)} \widetilde{V}_{m-k}\right]. \tag{3.4}$$

Proof. The initialization phase leads to the relation $\widetilde{Z}_m = \widetilde{M}_{\widetilde{V}_m}^{(0)} \widetilde{V}_m$. We suppose that at the end of the $(i-1)$th cycle the following relation holds: $\widetilde{Z}_m = \widetilde{M}_{\widetilde{V}_m}^{(i-1)} \widetilde{V}_m$. At step 9b of the ith cycle, $\widetilde{Z}_k$ is defined as

$$\widetilde{Z}_k = \widetilde{Z}_m \widetilde{Q}_{m\times k} = \widetilde{M}_{\widetilde{V}_m}^{(i-1)} \widetilde{V}_m \widetilde{Q}_{m\times k} = \widetilde{M}_{\widetilde{V}_m}^{(i-1)} \widetilde{V}_k.$$

The proof is then complete since $\widetilde{Z}_{m-k} = [\widetilde{\mathcal{M}}_{k+1}^{(i)}(\widetilde{v}_{k+1}), \ldots, \widetilde{\mathcal{M}}_m^{(i)}(\widetilde{v}_m)] = \widetilde{M}_{\widetilde{V}_{m-k}}^{(i)} \widetilde{V}_{m-k}$ at the end of step 10b. ∎

The next lemma details a relation between Z_m and W_m that is satisfied in the FGCRO-DR method.

LEMMA 3.3. *At the end of the* i*th cycle of the FGCRO-DR method,* Z_m *and* W_m *satisfy*

$$Z_m = M_{W_m}^{(i)} W_m = \left[M_{W_m}^{(i-1)} W_k,\ M_{W_{m-k}}^{(i)} W_{m-k}\right]. \tag{3.5}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Proof. The initialization phase leads to the relation $Z_m = M_{W_m}^{(0)} W_m$. We suppose that at the end of the $(i-1)$th cycle the following relation holds: $Z_m = M_{W_m}^{(i-1)} W_m$. At step 9a of the ith cycle, Z_k is defined as

$$\begin{aligned} Z_k &= Z_m P_k R^{-1} \\ &= M_{W_m}^{(i-1)} W_m P_k R^{-1} \\ &= M_{W_m}^{(i-1)} W_k. \end{aligned}$$

The proof is then complete since $Z_{m-k} = [\mathcal{M}_{k+1}^{(i)}(w_{k+1}), \ldots, \mathcal{M}_m^{(i)}(w_m)] = M_{W_{m-k}}^{(i)} W_{m-k}$ at the end of step 11a. ∎

Lemmas 3.2 and 3.3 show that $\widetilde{Z}_m$, $\widetilde{V}_m$, Z_m, and W_m satisfy relations that will play a central role in section 3.3. We investigate next the relation between Z_m and V_m.

LEMMA 3.4. *At the end of the ith cycle of the FGCRO-DR method, Z_m and V_m satisfy*

$$[AZ_k, Z_{m-k}] = \left[V_k,\ M_{V_{m-k}}^{(i)} V_{m-k}\right]. \tag{3.6}$$

Proof. We use the relation $AZ_k = V_k$ satisfied in the FGCRO-DR method shown in Proposition 2. The proof is then complete since $Z_{m-k} = [\mathcal{M}_{k+1}^{(i)}(v_{k+1}), \ldots, \mathcal{M}_m^{(i)}(v_m)] = M_{V_{m-k}}^{(i)} V_{m-k}$ at the end of step 11a. ∎

We conclude this section by presenting a technical lemma related to the FGMRES-DR method.

LEMMA 3.5. *During the ith cycle of the FGMRES-DR method, $\widetilde{v}_{k+1}$ satisfies the relation*

$$\widetilde{v}_{k+1} = \widetilde{\bar{v}}_{k+1} / ||\widetilde{\bar{v}}_{k+1}|| \quad \text{with} \quad \widetilde{\bar{v}}_{k+1} = \Pi_{[\widetilde{Y}_k]^\perp} \widetilde{r}_0^{(i-1)}, \tag{3.7}$$

where $\widetilde{r}_0^{(i-1)} = b - A\widetilde{x}_0^{(i-1)}$ denotes the residual obtained at the end of the $(i-1)$th cycle.

Proof. Using Proposition 1 and relation (2.8) we obtain

$$\begin{aligned} \widetilde{\bar{v}}_{k+1} &= \widetilde{V}_{m+1}\bar{\rho} = \widetilde{r}_0^{(i-1)} - \widetilde{V}_{m+1} \begin{bmatrix} \widetilde{Q}_{m\times k} \\ 0_{1\times k} \end{bmatrix} \begin{bmatrix} \widetilde{Q}_{m\times k} \\ 0_{1\times k} \end{bmatrix}^H \widetilde{V}_{m+1}^H \widetilde{r}_0^{(i-1)}, \\ \widetilde{\bar{v}}_{k+1} &= \widetilde{V}_{m+1}\bar{\rho} = \widetilde{r}_0^{(i-1)} - \widetilde{V}_m \widetilde{Q}_{m\times k} (\widetilde{V}_m \widetilde{Q}_{m\times k})^H \widetilde{r}_0^{(i-1)}. \end{aligned}$$

Since $\widetilde{V}_m \widetilde{Q}_{m\times k}$ has orthonormal columns, this last expression now becomes

$$\bar{v}_{k+1} = \Pi_{[\widetilde{V}_m \widetilde{Q}_{m\times k}]^\perp} \widetilde{r}_0^{(i-1)}.$$

Because $\widetilde{Q}_{m\times k}$ is the orthogonal factor of the QR decomposition of $\widetilde{P}_k$, we obtain

$$\text{range}(\widetilde{V}_m \widetilde{P}_k) = \text{range}(\widetilde{V}_m \widetilde{Q}_{m\times k}).$$

Since from Lemma 2.3 $\widetilde{Y}_k = \widetilde{V}_m \widetilde{P}_k$, the proof is then complete. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

3.3. Analysis of the FGMRES-DR and FGCRO-DR methods. Lemma 3.3 has already described an important property satisfied by W_m in the FGCRO-DR method proposed in Algorithm 1. We will analyze further the relation between the FGMRES-DR and FGCRO-DR methods. The next theorem states that the two flexible methods generate the same iterates in exact arithmetic under some conditions involving notably two vectors.

THEOREM 3.6. *We denote by* $r_0^{(i)} = b - Ax_0^{(i)}$ *the residual obtained at the end of the* i*th cycle of the FGCRO-DR method (see step* 14 *of Algorithm* 1*). We suppose that Definition* 3.1 *holds and that the same equivalent preconditioning matrix is obtained at the end of the* i*th cycle of both the FGCRO-DR and FGMRES-DR algorithms, i.e.,* $M_{W_m}^{(i)} = \widetilde{M}_{\widetilde{V}_m}^{(i)}$. *Under this assumption the harmonic Ritz vectors* $\widetilde{Y}_k$ *and* Y_k *can be chosen equal during the* $(i+1)$*th cycle. If in addition there exists a real-valued positive coefficient* η_{i+1} *such that*

$$\Pi_{[Y_k, r_0^{(i)}/\|r_0^{(i)}\|]^\perp} A\mathcal{M}_{k+1}^{(i+1)} \left(\Pi_{Y_k^\perp} r_0^{(i)}/\|\Pi_{Y_k^\perp} r_0^{(i)}\|\right) = \eta_{i+1}\ \Pi_{[Y_k, r_0^{(i)}/\|r_0^{(i)}\|]^\perp} A\mathcal{M}_{k+1}^{(i+1)} \left(r_0^{(i)}/\|r_0^{(i)}\|\right) \tag{3.8}$$

in the FGCRO-DR algorithm, then both algorithms generate the same iterates in exact arithmetic and

$$\mathrm{range}(V_{m+1}) = \mathrm{range}(\widetilde{V}_{m+1}), \tag{3.9}$$

$$\mathrm{range}(Z_m) = \mathrm{range}(\widetilde{Z}_m), \tag{3.10}$$

with

$$V_{m+1} = [\widetilde{V}_{k+1}\widehat{Q}, v_{k+2}, \ldots, v_{m+1}], \qquad \widetilde{V}_{m+1} = [\widetilde{V}_{k+1}, v_{k+2}, \ldots, v_{m+1}], \tag{3.11}$$

$$Z_m = [\widetilde{Z}_{k+1}\widehat{X}, z_{k+2}, \ldots, z_m], \qquad \widetilde{Z}_m = [\widetilde{Z}_{k+1}, z_{k+2}, \ldots, z_m], \tag{3.12}$$

where $\widehat{Q} \in \mathbb{C}^{(k+1)\times(k+1)}$ *is a unitary matrix and* $\widehat{X} \in \mathbb{C}^{(k+1)\times(k+1)}$ *is a nonsingular triangular matrix.*

Proof. The whole proof is performed in three parts assuming that we analyze the $(i+1)$th cycle of each algorithm. Suppose that at the beginning of the $(i+1)$th cycle (step 4) there exist a unitary matrix $\widehat{Q} \in \mathbb{C}^{(k+1)\times(k+1)}$ and a nonsingular matrix $\widehat{X} \in \mathbb{C}^{(k+1)\times(k+1)}$ such that the following relations hold:

$$V_{k+1} = \widetilde{V}_{k+1}\widehat{Q}, \tag{3.13}$$

$$Z_{k+1} = \widetilde{Z}_{k+1}\widehat{X}, \tag{3.14}$$

$$[v_{k+2}, \ldots, v_{m+1}] = [\widetilde{v}_{k+2}, \ldots, \widetilde{v}_{m+1}], \tag{3.15}$$

$$[z_{k+2}, \ldots, z_m] = [\widetilde{z}_{k+2}, \ldots, \widetilde{z}_m]. \tag{3.16}$$

We will then prove the existence of a unitary matrix $\widehat{Q}' \in \mathbb{C}^{(k+1)\times(k+1)}$ and of a nonsingular matrix $\widehat{X}' \in \mathbb{C}^{(k+1)\times(k+1)}$ such that at the end of the $(i+1)$th cycle

$$V_{k+1} = \widetilde{V}_{k+1}\widehat{Q}', \tag{3.17}$$

$$Z_{k+1} = \widetilde{Z}_{k+1}\widehat{X}', \tag{3.18}$$

$$[v_{k+2}, \ldots, v_{m+1}] = [\widetilde{v}_{k+2}, \ldots, \widetilde{v}_{m+1}], \tag{3.19}$$

$$[z_{k+2}, \ldots, z_m] = [\widetilde{z}_{k+2}, \ldots, \widetilde{z}_m]. \tag{3.20}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Regarding FGCRO-DR we assume that at the beginning of the $(i+1)$th cycle (step 4)

$$\text{range}(W_m) = \text{range}(V_m). \tag{3.21}$$

We will also prove that relation (3.21) holds at the end of the $(i+1)$th cycle. Note that relations (3.9), (3.10), and (3.21) are obviously satisfied before the first cycle, because steps 1 to 3 are identical in both algorithms, yielding $V_{m+1} = \widetilde{V}_{m+1}$, $Z_m = \widetilde{Z}_m$, and $W_m = V_m$. Finally a consequence of (3.13), (3.15), (3.14), and (3.16) is that the residuals of the linear system $Ax = b$ in both algorithms are equal at the beginning of the $(i+1)$th cycle, i.e., $r_0^{(i)} = \widetilde{r}_0^{(i)}$. We will denote by r_0 this residual for ease of notation.

Part I: *Steps* 5a *and* 5b. In this part, we prove that we can choose $\widetilde{Y}_k = Y_k$ with $Y_k = W_m P_k = \widetilde{V}_m \widetilde{P}_k$.

FGCRO-DR. Let $y_j = W_m p_j$ be the jth column of Y_k. Since y_j is a harmonic Ritz vector of $AZ_m W_m^{\dagger}$ with respect to range(W_m), the following relation holds (see Definition 2.1):

$$Z_m^H A^H \left(AZ_m p_j - \theta_j W_m p_j\right) = 0. \tag{3.22}$$

Due to (3.14) and (3.16) there exists a nonsingular matrix $X \in \mathbb{C}^{m\times m}$ that relates Z_m and $\widetilde{Z}_m$:

$$Z_m = \widetilde{Z}_m X. \tag{3.23}$$

Using (3.23), the harmonic Ritz relation (3.22) now becomes

$$X^H \widetilde{Z}_m^H A^H \left(A\widetilde{Z}_m X p_j - \theta_j W_m p_j\right) = 0.$$

From Lemma 3.3 and relation (3.23) we deduce

$$X^H \widetilde{Z}_m^H A^H \left(A\widetilde{Z}_m X p_j - \theta_j M_{W_m}^{(i)^{-1}} Z_m p_j\right) = 0,$$
$$X^H \widetilde{Z}_m^H A^H \left(A\widetilde{Z}_m X p_j - \theta_j \widetilde{M}_{\widetilde{V}_m}^{(i)^{-1}} \widetilde{Z}_m X p_j\right) = 0,$$

where we have used explicitly the assumption on the equivalent preconditioning matrix obtained at the end of the ith cycle, i.e., $M_{W_m}^{(i)} = \widetilde{M}_{\widetilde{V}_m}^{(i)}$. Next, the application of Lemma 3.2 leads to

$$X^H \widetilde{Z}_m^H A^H \left(A\widetilde{Z}_m \widetilde{V}_m^H \widetilde{V}_m X p_j - \theta_j \widetilde{V}_m X p_j\right) = 0. \tag{3.24}$$

Since X is nonsingular the last equality proves that $\widetilde{V}_m X p_j$ is a harmonic Ritz vector of $A\widetilde{Z}_m \widetilde{V}_m^H$ with respect to range$(\widetilde{V}_m)$ associated to the Ritz value θ_j. From relations (3.22) and (3.24) we deduce that the harmonic Ritz vectors can be chosen to be equal and correspond to the same harmonic Ritz values. In this case they notably satisfy the following equality:

$$\forall j \in \{1, \ldots, k\}, \quad \widetilde{V}_m X p_j = W_m p_j, \quad \text{i.e.,} \quad \widetilde{p}_j = X p_j. \tag{3.25}$$

We will then denote by $Y = \widetilde{Y}_k = Y_k$ the k harmonic Ritz vectors computed in either FGCRO-DR or FGMRES-DR. We assume that the harmonic Ritz values θ_j $(1 \le j \le k)$ are nonzero.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Part IIa: *Steps* 6a–10a, 6b–10b. We show that at the end of steps 10a and 10b the following relations hold: $\text{range}(V_{k+1}) = \text{range}(\widetilde{V}_{k+1}) = \text{range}([Y, r_0^{(i)}/\|r_0^{(i)}\|])$. This result will help us to prove the existence of the matrix $\widehat{Q}'$ introduced in relation (3.17).

FGCRO-DR. Since $AZ_mP_k = V_kR$ (Proposition 2), we deduce from Lemma 2.3

$$\left[V_k, r_0^{(i)}/\|r_0^{(i)}\|\right] = \left[Y, r_0^{(i)}/\|r_0^{(i)}\|\right] \begin{bmatrix} \text{diag}(\theta_1, \dots \theta_k)R^{-1} & 0_{k\times 1} \\ \|r_0^{(i)}\|\beta_{1\times k}R^{-1} & 1 \end{bmatrix}. \tag{3.26}$$

This relation leads to the following result:

$$\text{range}(V_{k+1}) = \text{range}\left(\left[Y, r_0^{(i)}/\|r_0^{(i)}\|\right]\right). \tag{3.27}$$

Similarly $W_{k+1} = [W_k, \frac{r_0^{(i)}}{\|r_0^{(i)}\|}]$, using $Y = W_mP_k$, can be written as

$$\begin{aligned} [W_k, r_0^{(i)}/\|r_0^{(i)}\|] &= \left[W_mP_kR^{-1}, \frac{r_0^{(i)}}{\|r_0^{(i)}\|}\right] \\ &= \left[YR^{-1}, r_0^{(i)}/\|r_0^{(i)}\|\right] \\ &= \left[Y, r_0^{(i)}/\|r_0^{(i)}\|\right] \begin{bmatrix} R^{-1} & 0_{k\times 1} \\ 0_{1\times k} & 1 \end{bmatrix}. \end{aligned} \tag{3.28}$$

From relations (3.28) and (3.27) we deduce that

$$\text{range}(W_{k+1}) = \text{range}(V_{k+1}). \tag{3.29}$$

This last result also proves that $\text{range}(W_m) = \text{range}(V_m)$ at the end of the cycle.

FGMRES-DR. In Proposition 1 we have shown that

$$\text{range}(\widetilde{V}_{k+1}) = \text{range}\left(\left[Y, r_0^{(i)}/\|r_0^{(i)}\|\right]\right). \tag{3.30}$$

Since both V_{k+1} and $\widetilde{V}_{k+1}$ have orthonormal columns, we deduce from (3.27) and (3.30) that there exists a unitary matrix $\widehat{Q}'$ such that

$$V_{k+1} = \widetilde{V}_{k+1}\widehat{Q}', \tag{3.31}$$

which proves the relation proposed in (3.17).

Part IIb: *Steps* 6a–10*a,* 6b–10b. We show that at the end of steps 10a and 10b the following relation holds: $\text{range}(Z_{k+1}) = \text{range}(\widetilde{Z}_{k+1})$. This result will help us to prove the existence of the matrix $\widehat{X}'$ introduced in relation (3.18).

FGCRO-DR. Concerning $Z_{k+1} = [Z_k, z_{k+1}]$, there exists a nonsingular matrix $M^{(i+1)}_{[W_k, r_0^{(i)}/\|r_0^{(i)}\|]} \in \mathbb{C}^{n\times n}$ (see Definition 3.1) such that

$$Z_{k+1} = M^{(i+1)}_{[W_k, r_0^{(i)}/\|r_0^{(i)}\|]} \left[W_k, r_0^{(i)}/\|r_0^{(i)}\|\right].$$

If $T \in \mathbb{C}^{(k+1)\times(k+1)}$ denotes the triangular matrix

$$T = \begin{bmatrix} R & 0_{k\times 1} \\ 0_{1\times k} & 1 \end{bmatrix}$$

due to relation (3.28), then $Z_{k+1}T$ can be written as

$$Z_{k+1}T = M^{(i+1)}_{[W_k, r_0^{(i)}/\|r_0^{(i)}\|]} \left[Y, r_0^{(i)}/\|r_0^{(i)}\|\right]. \tag{3.32}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

FGMRES-DR. Similarly, from Lemma 3.2, $\widetilde{Z}_{k+1}$ can be expressed as

$$\widetilde{Z}_{k+1} = \widetilde{M}^{(i+1)}_{\widetilde{V}_{k+1}} \widetilde{V}_{k+1},$$

where $\widetilde{M}^{(i+1)}_{\widetilde{V}_{k+1}} \in \mathbb{C}^{n\times n}$ is nonsingular (see Definition 3.1). If $\widetilde{T} \in \mathbb{C}^{(k+1)\times(k+1)}$ denotes the triangular matrix

$$\widetilde{T} = \widetilde{R} \begin{bmatrix} I_k & 0_{k\times 1} \\ 0_{1\times k} & 1/\|r_0^{(i)}\| \end{bmatrix},$$

$\widetilde{Z}_{k+1}\widetilde{T}$ can be expressed as

$$\widetilde{Z}_{k+1}\widetilde{T} = \widetilde{M}^{(i+1)}_{\widetilde{V}_{k+1}} \left[Y, r_0^{(i)}/\|r_0^{(i)}\| \right] \tag{3.33}$$

thanks to relation (2.15). Relations (3.32) and (3.33) characterize $Z_{k+1}T$ and $\widetilde{Z}_{k+1}\widetilde{T}$ with respect to $[Y, r_0^{(i)}/\|r_0^{(i)}\|]$. We can further improve this result by showing the following equality:

$$M^{(i+1)}_{[W_k, r_0^{(i)}/\|r_0^{(i)}\|]} \left[Y, r_0^{(i)}/\|r_0^{(i)}\| \right] = \widetilde{M}^{(i+1)}_{\widetilde{V}_{k+1}} \left[Y, r_0^{(i)}/\|r_0^{(i)}\| \right]. \tag{3.34}$$

Lemma 3.3 and Lemma 3.2, respectively, give us two useful relations for $M^{(i+1)}_{[W_k, r_0^{(i)}/\|r_0^{(i)}\|]}$ $[Y, r_0^{(i)}/\|r_0^{(i)}\|]$ and $\widetilde{M}^{(i+1)}_{\widetilde{V}_{k+1}}[Y, r_0^{(i)}/\|r_0^{(i)}\|]$, i.e.,

$$M^{(i+1)}_{[W_k, r_0^{(i)}/\|r_0^{(i)}\|]} \left[Y, r_0^{(i)}/\|r_0^{(i)}\| \right] = \left[M^{(i)}_{W_m} Y,\ \mathcal{M}^{(i+1)}_{k+1} \left(r_0^{(i)}/\|r_0^{(i)}\| \right) \right], \tag{3.35}$$

$$\widetilde{M}^{(i+1)}_{\widetilde{V}_{k+1}} \left[Y, r_0^{(i)}/\|r_0^{(i)}\| \right] = \left[\widetilde{M}^{(i)}_{\widetilde{V}_m} Y,\ \widetilde{\mathcal{M}}^{(i+1)}_{k+1} \left(r_0^{(i)}/\|r_0^{(i)}\| \right) \right]. \tag{3.36}$$

Using the assumption on the equivalent preconditioning matrix obtained at the end of the ith cycle, i.e., $M^{(i)}_{W_m} = \widetilde{M}^{(i)}_{\widetilde{V}_m}$, we have

$$M^{(i)}_{W_m} Y = \widetilde{M}^{(i)}_{\widetilde{V}_m} Y. \tag{3.37}$$

The fact that identical (possibly nonlinear) preconditioning operators are used in steps 10a and 10b of Algorithm 1 (see relation (3.1)) allows us to write

$$\mathcal{M}^{(i+1)}_{k+1} \left(r_0^{(i)}/\|r_0^{(i)}\| \right) = \widetilde{\mathcal{M}}^{(i+1)}_{k+1} \left(r_0^{(i)}/\|r_0^{(i)}\| \right). \tag{3.38}$$

Relations (3.37) and (3.38) finally show the relation (3.34). Consequently from relations (3.32), (3.33), and (3.34) we deduce that there exists a nonsingular matrix $\widehat{X}' \in \mathbb{C}^{(k+1)\times(k+1)}$ such that

$$Z_{k+1} = \widetilde{Z}_{k+1}\widehat{X}'. \tag{3.39}$$

This proves the relation proposed in (3.18). Since T and $\widetilde{T}$ are both triangular, we note that $\widehat{X}' = \widetilde{T}T^{-1}$ is also triangular.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Part IIIa: *Steps* 10a *and* 10b. We first show that $\widetilde{v}_{k+2} = v_{k+2}$ by expressing these two quantities as a function of $r_0^{(i)}$ and Y.

FGCRO-DR. The Arnoldi relation (step 10a) yields $v_{k+2} = \bar{v}_{k+2}/||\bar{v}_{k+2}||$, where $\bar{v}_{k+2} = (I_n - v_{k+1}v_{k+1}^H)(I_n - V_kV_k^H)A\mathcal{M}_{k+1}^{(i+1)}(r_0^{(i)}/\|r_0^{(i)}\|)$. Since from Proposition 2 $V_k^H r_0^{(i)} = 0$ in the $(i+1)$th cycle, $(I_n - v_{k+1}v_{k+1}^H)$ and $(I_n - V_kV_k^H)$ commute, and from Part IIa of the proof, the following expression can be derived:

$$\bar{v}_{k+2} = \Pi_{V_{k+1}^\perp} A\mathcal{M}_{k+1}^{(i+1)}(r_0^{(i)}/\|r_0^{(i)}\|) = \Pi_{[Y, r_0^{(i)}/\|r_0^{(i)}\|]^\perp} A\mathcal{M}_{k+1}^{(i+1)}(r_0^{(i)}/\|r_0^{(i)}\|). \tag{3.40}$$

FGMRES-DR. The following expression for $\widetilde{v}_{k+2} = \widetilde{\bar{v}}_{k+2}/||\widetilde{\bar{v}}_{k+2}||$ is obtained using Lemma 3.5:

$$\widetilde{\bar{v}}_{k+2} = (I_n - \widetilde{V}_{k+1}\widetilde{V}_{k+1}^H)A\mathcal{M}_{k+1}^{(i+1)}(\widetilde{v}_{k+1}) = \Pi_{[Y, r_0^{(i)}/\|r_0^{(i)}\|]^\perp} A\mathcal{M}_{k+1}^{(i+1)}(\Pi_{Y^\perp} r_0^{(i)}/\|\Pi_{Y^\perp} r_0^{(i)}\|). \tag{3.41}$$

Due to the assumption (3.8) of Theorem 3.6 we deduce from (3.40) and (3.41) that $\bar{v}_{k+2} = \eta_{i+1}\ \widetilde{\bar{v}}_{k+2}$ with η_{i+1} positive, and therefore $v_{k+2} = \widetilde{v}_{k+2}$.

Part IIIb: *Steps* 10a *and* 10b. In this part we continue the analysis of the Arnoldi procedure with flexible preconditioning and show that $v_{k+2+j} = \widetilde{v}_{k+2+j}$ for $j = 1, \ldots, m-k-1$.

For the case $j = 1$, we introduce $\bar{v}_{k+3}$ and $\widetilde{\bar{v}}_{k+3}$ such that $v_{k+3} = \bar{v}_{k+3}/||\bar{v}_{k+3}||$ and $\widetilde{v}_{k+3} = \widetilde{\bar{v}}_{k+3}/||\widetilde{\bar{v}}_{k+3}||$. The application of the Arnoldi procedure in both algorithms leads to

$$\begin{aligned}
\bar{v}_{k+3} &= (I_n - v_{k+2}v_{k+2}^H)(I_n - V_{k+1}V_{k+1}^H)\ A\mathcal{M}_{k+2}^{(i+1)}(\bar{v}_{k+2}),\\
\widetilde{\bar{v}}_{k+3} &= (I_n - \widetilde{v}_{k+2}\widetilde{v}_{k+2}^H)(I_n - \widetilde{V}_{k+1}\widetilde{V}_{k+1}^H)\ A\mathcal{M}_{k+2}^{(i+1)}(\widetilde{\bar{v}}_{k+2}).
\end{aligned}$$

Thus from Parts II and IIIa of the proof we obtain that v_{k+3} and $\widetilde{v}_{k+3}$ are equal. The proof can then be completed by induction.

Results from Parts II and III justify relation (3.19), i.e., $[v_{k+2}, \ldots, v_{m+1}] = [\widetilde{v}_{k+2}, \ldots, \widetilde{v}_{m+1}]$. Consequently from Lemma 3.2, Lemma 3.4, and relation (3.1) we deduce relation (3.20). This finally shows the main relations (3.9) and (3.10) of Theorem 3.6 that are satisfied at the end of the $(i+1)$th cycle. ∎

3.3.1. First consequence of Theorem 3.6.

COROLLARY 3.7. *If the same flexible preconditioning operators are used in both Arnoldi procedures (steps* 10a *and* 10b *of Algorithm* 1*) and if at each cycle i there exists a real-valued positive coefficient η_i such that*

$$\begin{aligned}
&\Pi_{[Y, r_0^{(i-1)}/\|r_0^{(i-1)}\|]^\perp} A\mathcal{M}_{k+1}^{(i)}(\Pi_{Y^\perp} r_0^{(i-1)}/\|\Pi_{Y^\perp} r_0^{(i-1)}\|)\\
&\quad = \eta_i\ \Pi_{[Y, r_0^{(i-1)}/\|r_0^{(i-1)}\|]^\perp} A\mathcal{M}_{k+1}^{(i)}(r_0^{(i-1)}/\|r_0^{(i-1)}\|),
\end{aligned}$$

or, equivalently (from relations (3.40) *and* (3.41)*), such that $\widetilde{\bar{v}}_{k+2} = \eta_i\ \bar{v}_{k+2}$, FGCRO-DR and FGMRES-DR are algebraically equivalent.*

Proof. We have already emphasized that $M_{W_m}^{(0)} = \widetilde{M}_{\widetilde{V}_m}^{(0)}$ in relation (3.3). In Theorem 3.6 we have analyzed the $(i+1)$th cycle of both algorithms assuming that $M_{W_m}^{(i)} = \widetilde{M}_{\widetilde{V}_m}^{(i)}$. First we have proved in Part IIb the relation (3.34), and second in Parts IIIa and IIIb that $[v_{k+2}, \ldots, v_m] = [\widetilde{v}_{k+2}, \ldots, \widetilde{v}_m]$ and $[z_{k+2}, \ldots, z_m] = [\widetilde{z}_{k+2}, \ldots, \widetilde{z}_m]$, respectively. Consequently the same equivalent preconditioner matrix is obtained at the end of the $(i+1)$th cycle, i.e., $M_{W_m}^{(i+1)}$ and $\widetilde{M}_{\widetilde{V}_m}^{(i+1)}$ can be chosen equal. We deduce that FGCRO-DR and FGMRES-DR are algebraically equivalent. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

3.3.2. About GCRO-DR and GMRES-DR. We propose a second consequence of Theorem 3.6 analyzed now with a fixed preconditioning matrix M.

COROLLARY 3.8. *When a fixed right preconditioner is used, the GCRO-DR and GMRES-DR methods sketched in Algorithm* 1 *are algebraically equivalent.*

Proof. We denote by M the fixed right preconditioning operator. A straightforward reformulation of Lemma 3.3 in this context leads to the relation $Z_m = MW_m$ in GCRO-DR. Exploiting now Lemma 2.3 allows us to derive the following relation, which holds during the $(i+1)$th cycle:

$$AMW_mP_k = AMY = [Y, r_0^{(i)}] \begin{bmatrix} \operatorname{diag}(\theta_1, \ldots, \theta_k) \\ \beta_{1\times k} \end{bmatrix}.$$

Thus

$$\Pi_{[Y, r_0^{(i)}]^\perp} AMY = 0. \tag{3.42}$$

Due to (3.42) and Part IIIa of the proof of Theorem 3.6 we deduce the following development:

$$\begin{aligned} \bar{v}_{k+2} &= \Pi_{[Y, r_0^{(i)}]^\perp} AM \left(r_0^{(i)} - YY^\dagger r_0^{(i)} \right), \\ \bar{v}_{k+2} &= \Pi_{[Y, r_0^{(i)}]^\perp} AM \Pi_{Y^\perp} r_0^{(i)}, \\ \bar{v}_{k+2} &= \widetilde{\bar{v}}_{k+2}. \end{aligned}$$

By induction it is possible to deduce the rest of the proof regarding $\bar{v}_{k+j}, j > 2$. Using $\operatorname{range}(\widetilde{V}_{k+1}) = \operatorname{range}(V_{k+1})$ obtained in Part IIa we deduce that

$$\operatorname{range}(\widetilde{V}_m) = \operatorname{range}(V_m) = \operatorname{range}(W_m). \tag{3.43}$$

A straightforward reformulation of Lemma 3.2 leads to the relation $\widetilde{Z}_m = M\widetilde{V}_m$ in GMRES-DR. From relation (3.43) we finally deduce that

$$\operatorname{range}(\widetilde{Z}_m) = \operatorname{range}(Z_m).$$

Consequently the minimization problem $\min \|r_0^{(i)} - AZ_m y\|$ leads to the same solution for both algorithms at each cycle: GCRO-DR and GMRES-DR sketched in Algorithm 1 are thus algebraically equivalent. ∎

3.3.3. A numerical illustration. In this section we intend to illustrate the results shown in sections 3.3.1 and 3.3.2 on a simple numerical example. We consider a real symmetric positive definite matrix $A = Q\ D\ Q^T$ of size 200 with Q orthonormal and D diagonal with entries ranging from 10^{-4} to 1. The spectrum of A contains eigenvalues of small magnitude,[1] and consequently the use of deflation techniques should improve the convergence rate of Krylov subspace methods if the harmonic Ritz values of smallest modulus are taken into account. In this experiment we consider a polynomial preconditioner represented by two iterations of unpreconditioned GMRES for the solution of $Ax = b$ with b given by $b = \frac{Ae}{\|Ae\|_2}$ ($e \in \mathbb{R}^{200}$ denoting the vector with all components equal to one) starting from a zero initial guess. Figure 3.1 shows the histories of convergence of various flexible methods minimizing

[1] The eigenvalues of A are logarithmically spaced $(10^{-4}, 10^{-3}, 10^{-2})$ and linearly distributed between 0.02 and 1 with step 1/200.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

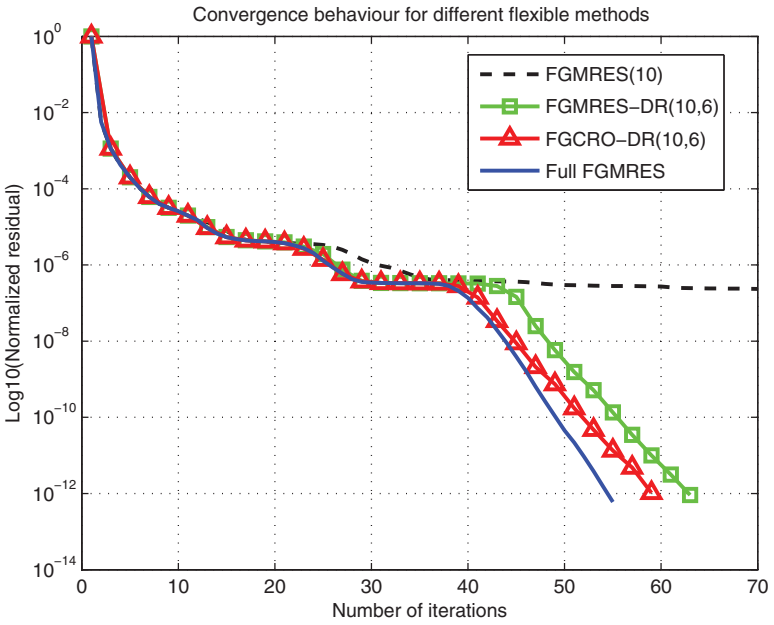


FIG. 3.1. *Convergence histories of different flexible methods applied to* $Ax = b$, *where* $A \in \mathbb{R}^{200\times 200}$ *is symmetric positive definite with some eigenvalues of small magnitude.*

TABLE 3.1

Scalar product $v_{k+2}^T \widetilde{v}_{k+2}$ *during the first five cycles of FGCRO-DR*(10,6) *when solving the linear system considered in section* 3.3.3 *using a variable preconditioner (two iterations of GMRES) and a fixed right preconditioner (diagonal preconditioning).*

Cycle index	1	2	3	4	5
Variable preconditioner	0.92	0.89	0.45	0.90	0.90
Fixed right preconditioner	1.00	1.00	1.00	1.00	1.00

over a subspace of same dimension, i.e., FGMRES(10), FGMRES-DR(10,6), FGCRO-DR(10,6), respectively, and full flexible GMRES with such a variable preconditioner. Flexible methods with deflated restarting are found to be efficient since they are close to the full flexible GMRES method in terms of performances. We also remark that the convergence histories of FGCRO-DR(10,6) and FGMRES-DR(10,6) are different. According to Corollary 3.7 we compute the scalar product of v_{k+2} and $\widetilde{v}_{k+2}$ (which are both vectors of unit norm) to determine the cosine of the angle between these two vectors. The values are reported in Table 3.1 for the first five cycles. With such a variable preconditioner it is found that the methods are not equivalent in the first cycle already since the collinearity condition between v_{k+2} and $\widetilde{v}_{k+2}$ is not fulfilled. The situation is similar during the following cycles, which explains why different convergence histories for FGMRES-DR(10,6) and FGCRO-DR(10,6) observed in Figure 3.1 are obtained in such a case. As expected from section 3.3.2, if a fixed right preconditioner is used, the convergence histories of GMRES-DR(10,6) and GCRO-DR(10,6) are found to be exactly the same (results not shown here). In such a case v_{k+2} and $\widetilde{v}_{k+2}$ fulfill the collinearity condition; this is confirmed in Table 3.1 when a diagonal preconditioning is used.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

4. Further features of FGCRO-DR(m, k). In this section we first compare FGCRO-DR(m, k) with FGMRES-DR(m, k) presented in Algorithm 1 from both a computational and a storage point of view. Then we detail how subspace recycling can be used in FGCRO-DR(m, k) when solving a sequence of linear systems.

4.1. Computational cost. We first analyze the computational cost related to the generalized eigenvalue problem to deduce harmonic Ritz information and then detail the total cost of the proposed method.

4.1.1. Harmonic Ritz information. The generalized eigenvalue problem (3.22) can also be written as

$$\bar{H}_m^H \ \bar{H}_m y = \theta \bar{H}_m^H V_{m+1}^H W_m y. \tag{4.1}$$

Since $W_m = [W_{k+1}, v_{k+2}, \ldots, v_m]$, $V_{m+1}^H W_m$ can be decomposed at the end of the cycle as

$$V_{m+1}^H W_m = \begin{bmatrix} V_{k+1}^H \ W_{k+1} & 0_{(k+1)\times(m-k-1)} \\ 0_{(m-k-1)\times(k+1)} & I_{m-k-1} \\ 0_{1\times(k+1)} & 0_{1\times(m-k-1)} \end{bmatrix}, \tag{4.2}$$

where the structure of the $(k+1)\times(k+1)$ block $V_{k+1}^H \ W_{k+1}$ is as follows:

$$V_{k+1}^H \ W_{k+1} = \begin{bmatrix} V_k^H \ W_k & V_k^H \ w_{k+1} \\ v_{k+1}^H \ W_k & v_{k+1}^H \ w_{k+1} \end{bmatrix} = \begin{bmatrix} V_k^H \ W_k & 0_{k\times 1} \\ v_{k+1}^H \ W_k & 1 \end{bmatrix}.$$

$V_k^H \ W_k$ is a $k \times k$ matrix that satisfies the following relation at the end of the ith cycle:

$$(V_k^H W_k)^{(i)} = Q^H \ (V_{m+1}^H W_m)^{(i-1)} \ P_k R^{-1},$$

where the superscript is related to the cycle index. Thus storing the $(m+1)\times m$ matrix $(V_{m+1}^H W_m)^{(i-1)}$ can be used to obtain $(V_k^H W_k)^{(i)}$ at a cost that is independent of n. Next we analyze how to compute efficiently $v_{k+1}^H \ W_k$ during the ith cycle. From relation (2.18) shown in Lemma 2.3 and Proposition 2, respectively, we deduce the relation

$$v_{k+1}^H V_k R = v_{k+1}^H W_k R \ \mathrm{diag}(\theta_1, \ldots, \theta_k) + v_{k+1}^H r_0^{(i-1)} \beta_{1\times k}. \tag{4.3}$$

Due to Proposition 2 and the definition of v_{k+1}, we have $v_{k+1}^H V_k = 0$. Thus we finally obtain that

$$v_{k+1}^H W_k = -\|(c - \bar{H}_m y^*)^{(i-1)}\|_2 \ \beta_{1\times k} \ (R \ \mathrm{diag}(\theta_1, \ldots, \theta_k))^{-1}, \tag{4.4}$$

where $\beta_{1\times k}$ is obtained from relation (2.19), which does only involve projected quantities. This allows us to deduce $v_{k+1}^H W_k$ at a cost independent of n. From this development we draw two important consequences from a computational point of view. First, $(V_{m+1}^H W_m)^{(i)}$ can be obtained recursively at a cost that is independent of the problem size n. Second, storing W_m (which would represent m additional vectors of size n) is not mandatory; only $V_{m+1}^H W_m$—matrix of size $(m+1)\times m$—is required.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

TABLE 4.1
Computational cost of a generic cycle of FGMRES-DR(m,k) and FGCRO-DR(m,k). C represents the total cost of FGCRO-DR(m,k) and corresponds to $C=(m-k)(op_M+op_A)+n(2(m+1)k+1+2mk+(m-k)(2m+2k+6))$.

Computation of	FGMRES-DR(m,k)	FGCRO-DR(m,k)
$V_m(:,1:k+1)$	$2n(m+1)(k+1)$	$2n(m+1)k+n$
$Z_m(:,1:k)$	$2nmk$	$2nmk$
$V_m(:,k+2:m+1)$	$(m-k)op_A+$ $n(m-k)(2m+2k+5)$	$(m-k)op_A+$ $n(m-k)(2m+2k+6)$
$Z_m(:,k+1:m)$	$(m-k)op_M$	$(m-k)op_M$
Total cost	$C+n\ (m+k+1)$	C

4.1.2. Cost of a cycle. We summarize in Table 4.1 the main computational costs associated with each generic cycle of FGMRES-DR(m,k) and FGCRO-DR(m,k). In FGCRO-DR(m,k), an Arnoldi method based on the modified Gram–Schmidt procedure has been assumed.[2] We have included only the costs proportional to the size of the original problem n which is supposed to be much greater than m and k. We denote by op_A and op_M the floating point operation counts for the matrix-vector product and the preconditioner application, respectively.

The generalized eigenvalue problem in FGCRO-DR(m,k) has been ignored in Table 4.1 since it can be performed at a cost independent of n as outlined in section 4.1.1. Furthermore the computation of c (required at step 12 of Algorithm 1) has not been included in Table 4.1 since in both methods it can be obtained at a cost independent of n (see Proposition 3 in [15] for FGMRES-DR). From Table 4.1 we deduce that FGCRO-DR(m,k) requires slightly fewer operations per cycle than FGMRES-DR(m,k).

4.2. Storage requirements. We consider only the storage proportional to the size of the original problem n. Similarly, as in FGMRES-DR(m,k) (see [15, section 3.2.2]), if the matrix multiplications $V_{m+1}Q$ and $Z_mP_kR^{-1}$ at steps 8a and 9a of Algorithm 1 are performed *in place* (i.e., overwriting V_k and Z_k, respectively), FGCRO-DR(m,k) requires only the storage of Z_m and V_{m+1}, which corresponds to $(2m+1)$ vectors of length n. The same storage cost is needed in FGMRES-DR(m,k) as detailed in [15].

4.3. Solution of sequence of linear systems. As advocated in [22], GCRO-DR(m,k) is suited for the solution of a sequence of slowly changing linear systems defined as $A^lx^l=b^l$ where both the matrix $A^l\in\mathbb{C}^{n\times n}$ and the right-hand side $b^l\in\mathbb{C}^n$ change from one system to the next, and the linear systems may typically not be available simultaneously. Next, we analyze how subspace recycling can be used in FGCRO-DR(m,k). We suppose that FGCRO-DR(m,k) has been applied for the solution of a given linear system (indexed by $s-1$) in this sequence and that appropriate subspaces to be recycled, Z_k^{s-1} and W_k^{s-1}, have been selected during a given cycle.

[2]In FGCRO-DR(m,k) (step 10a of Algorithm 1) the action of $(I_n-V_kV_k^H)$ requires $\sum_{j=k+1}^{m}(4nk+n)$ operations, the Arnoldi method based on modified Gram–Schmidt requires $\sum_{j=k+1}^{m}\sum_{i=k+1}^{j}(4n)$ operations, whereas norm computation and normalization cost $\sum_{j=k+1}^{m}(3n)$ operations. In FGMRES-DR(m,k) (step 10b of Algorithm 1) the Arnoldi method based on modified Gram–Schmidt requires $\sum_{j=k+1}^{m}\sum_{i=1}^{j}(4n)$ operations due to maintaining orthogonality to V_k, whereas norm computation and normalization cost $\sum_{j=k+1}^{m}(3n)$ operations.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

As explained in Proposition 2, the relations $A^{s-1}Z_k^{s-1} = V_k^{s-1}$ with ${V_k^{s-1}}^H V_k^{s-1} = I_k$ and range(W_k^{s-1}) = range(V_k^{s-1}) are then supposed to hold. Proposition 3 details how to consider subspace recycling in the initial phase of FGCRO-DR(m,k), when solving the new linear system $A^s x^s = b^s$ with x_0 as an initial guess.

PROPOSITION 3. *Suppose that Z_k^{s-1} and W_k^{s-1} are defined from solving a previous linear system $A^{s-1}x^{s-1} = b^{s-1}$ with FGCRO-DR(m,k) and that $A^s x^s = b^s$ is the new linear system to be solved. In the initial phase of FGCRO-DR with subspace recycling, the relation $A^s Z_k^s = V_k^s$ with ${V_k^s}^H V_k^s = I_k$ holds with matrices $V_k^s, Z_k^s \in \mathbb{C}^{n\times k}$ defined as*

$$\begin{aligned} V_k^s &= Q, \\ Z_k^s &= Z_k^{s-1}R^{-1} \end{aligned}$$

with $QR = A^s Z_k^{s-1}$, where $Q \in \mathbb{C}^{n\times k}$ has orthonormal columns and $R \in \mathbb{C}^{k\times k}$ is upper triangular. In addition we define $W_k^s \in \mathbb{C}^{n\times k}$ as $W_k^s = W_k^{s-1}R^{-1}$.

Proof. By using information related to the reduced QR factorization of $A^s Z_k^{s-1}$ and the relation $A^{s-1}Z_k^{s-1} = V_k^{s-1}$, respectively, we easily obtain

$$\begin{aligned} A^s Z_k^s &= A^s Z_k^{s-1}R^{-1} = Q \\ &= V_k^s. \end{aligned}$$

Since Q has orthonormal columns, V_k^s satisfies ${V_k^s}^H V_k^s = I_k$. Finally $W_k^s = W_k^{s-1}R^{-1}$ is imposed to make sure that the relation shown in Lemma 3.3 will hold at the end of the initial phase of FGCRO-DR(m,k) with subspace recycling. ∎

In the case of a sequence where only the right-hand sides are changing, we note that the reduced QR factorization (step 3 in Algorithm 2) is not required. The complete construction of the initial generation of subspaces V_{m+1}^s, Z_m^s, W_m^s is sketched in Algorithm 2. Once V_{m+1}^s, Z_m^s, and W_m^s have been obtained, the main cycle of FGCRO-DR(m,k) (lines 4 to 15 of Algorithm 1) can be applied straightforwardly.

Algorithm 2. Initial generation of V_{m+1}^s, Z_m^s, and W_m^s when subspace recycling is used to solve $A^s x^s = b^s$.

1: Suppose that V_k^{s-1}, Z_k^{s-1} and W_k^{s-1} are defined from solving a previous linear system $A^{s-1}x^{s-1} = b^{s-1}$ and satisfy $A^{s-1}Z_k^{s-1} = V_k^{s-1}$ with ${V_k^{s-1}}^H V_k^{s-1} = I_k$ and range(W_k^{s-1}) = range(V_k^{s-1})
2: $r_0 = b^s - A^s x_0$
3: $Q\ R = A^s Z_k^{s-1}$
4: $V_k^s = Q$
5: $Z_k^s = Z_k^{s-1}R^{-1}$
6: $W_k^s = W_k^{s-1}R^{-1}$
7: $x_0^{(0)} = x_0 + Z_k^s {V_k^s}^H r_0$
8: $r_0^{(0)} = r_0 - V_k^s {V_k^s}^H r_0$
9: Apply $m-k$ flexible preconditioned Arnoldi steps with $(I_n - V_k^s {V_k^s}^H)A^s$ and $v_{k+1}^s = r_0^{(0)}/\|r_0^{(0)}\|$ such that $(I_n - V_k^s {V_k^s}^H)A^s \left[z_{k+1}^s, \ldots, z_m^s\right] = \left[v_{k+1}^s, \ldots, v_{m+1}^s\right] \bar{H}_{m-k}$ with $z_j^s = \mathcal{M}_j^{(i)}(v_j^s)$
10: $d^* = \arg\min_{d\in Z_m^s} \|r_0^{(0)} - A^s d\|$, $x_0^{(1)} = x_0^{(0)} + d^*$, $r_0^{(1)} = b^s - A^s x_0^{(1)}$
11: $W_m^s = \left[W_k^s \quad V_m^s(1:n, k+1:m)\right]$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

TABLE 4.2

Solution of a d-dimensional elliptic partial differential equation problem on a 16^d grid with homogeneous Dirichlet boundary conditions ($d = 2, \ldots, 5$). Shown are the total number of matrix-vector products ($\#Mvp$) required to solve a sequence of twelve linear systems with different flexible methods. The variable preconditioner is based on four iterations of unpreconditioned GMRES. The stopping criterion corresponds to a reduction of six orders of magnitude of the normalized residual in the Euclidean norm. Harmonic Ritz values of smallest modulus have been considered when deflating.

Grid	16^2	16^3	16^4	16^5
Problem size (n)	(225)	(3375)	(50625)	(759375)
Method	$\#Mvp$	$\#Mvp$	$\#Mvp$	$\#Mvp$
FGMRES(20)	972	1176	1272	1128
FGMRES-DR(20,10)	732	948	1020	876
FGCRO-DR(20,10) (no recycling)	732	948	1020	876
FGCRO-DR(20,10) (with recycling)	457	541	547	529

Subspace recycling can thus be easily used in FGCRO-DR(m, k) to solve sequences of linear systems.

4.3.1. A numerical illustration. As a numerical illustration we consider sequences of linear systems arising from the finite difference discretization of multi-dimensional elliptic partial differential equations (isotropic Laplace operator) posed on the $[0, 1]^d$ hypercube with homogeneous Dirichlet boundary conditions. These sequences correspond to situations where only the right-hand sides are changing for a given dimension d. An efficient solution method is of primary interest in certain applications related to, e.g., financial engineering, molecular biology, or quantum dynamics [5, 6]. In the numerical experiments reported here (performed in MATLAB) we have used second order finite difference discretization schemes leading to sparse matrices with at most $2d + 1$ nonzero elements per row. We analyze the performances of various flexible methods used with four iterations of unpreconditioned GMRES as a preconditioner. This polynomial preconditioner is a variable nonlinear function which thus requires a flexible Krylov subspace method as an outer method [28]. Table 4.2 collects the number of matrix-vector products of some flexible methods minimizing over a subspace of the same dimension, i.e., FGMRES(20), FGMRES-DR(20,10), FGCRO-DR(20,10), and FGCRO-DR(20,10) with subspace recycling, respectively. Using deflation helps to improve the convergence rate of FGMRES in this application since a reduction of approximately 20% to 25% in terms of matrix-vector products is obtained for FGMRES-DR(20,10) independently of the dimension d. FGCRO-DR(20,10) leads to numbers of matrix-vector products which are similar to FGMRES-DR(20,10) although the convergence histories are found to be different. Finally, using both deflation and recycling in FGCRO-DR leads to a significant decrease in terms of matrix-vector products. A reduction in the range of 40% to 45% is indeed obtained versus another flexible Krylov subspace method with deflated restarting (FGMRES-DR(m, k)). This can be considered as a primary advantage over FGMRES-DR(m, k) since FGMRES-DR(m, k) does not allow subspace recycling. It nicely extends to the flexible setting the advantage of GCRO-DR versus GMRES-DR previously illustrated in [22]. We note that the resulting method is factorization free and mostly relies on matrix-vector products, a nice feature if distributed memory platforms are targeted to address numerical problems of larger size in higher dimension.

5. Conclusion and perspectives. In this paper we have studied a new minimum residual norm subspace method with deflated restarting that allows flexible

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

preconditioning based on the GCRO subspace method. The resulting method, named FGCRO-DR, has been presented together with FGMRES-DR, a recently proposed algorithm of the same family but based on the GMRES subspace method. A theoretical comparison analysis of both algorithms has been performed in section 3, where Theorem 3.6—the main result of this paper—proves the algebraic equivalence of FGMRES-DR and FGCRO-DR if a certain collinearity condition holds at each cycle. Corollary 3.8 has also proved that GMRES-DR and GCRO-DR are algebraically equivalent when a fixed right preconditioner is used. Furthermore we have carefully analyzed the computational cost of a given cycle of FGCRO-DR and have shown that FGCRO-DR is nearly as expensive as FGMRES-DR in terms of operations. FGCRO-DR offers the additional advantage of being suitable for the solution of sequences of slowly changing linear systems (where both the matrix and right-hand side can change) through subspace recycling.

In [8] variants of FGCRO-DR have been proposed which only differ in the formulation of the projected generalized eigenvalue problem. In future work we plan to investigate the numerical properties of these variants on realistic problems of large size for both single and multiple left- or right-hand side situations. Of interest are applications related to, e.g., steady or unsteady simulations of nonlinear equations [7] or stochastic finite element methods [12, 33] in three dimensions where variable preconditioning using approximate solvers has to be usually considered. We also note that when all right-hand sides are available simultaneously and when the matrix is fixed, block subspace methods may be also suitable. Thus a perspective could be to propose a block variant of FGCRO-DR.

Acknowledgments. We would like to thank the referees and the associate editor for their careful readings and valuable suggestions that helped us to improve our manuscript significantly. We also thank Iain S. Duff and Xavier Pinel for fruitful discussions and comments. The first author would like to acknowledge the warm welcome he received at CERFACS, in the Parallel Algorithms Team, during his sabbatical leave where the initial work was completed.

REFERENCES

[1] O. Axelsson and P. S. Vassilevski, *A black box generalized conjugate gradient solver with inner iterations and variable-step preconditioning*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 625–644.

[2] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, UK, 1994.

[3] J. Baglama, D. Calvetti, G. H. Golub, and L. Reichel, *Adaptively preconditioned GMRES algorithms*, SIAM J. Sci. Comput., 20 (1998), pp. 243–269.

[4] A. H. Baker, E. R. Jessup, and T. Manteuffel, *A technique for accelerating the convergence of restarted GMRES*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 962–984.

[5] G. Beylkin and M. J. Mohlenkamp, *Algorithms for numerical analysis in high dimensions*, SIAM J. Sci. Comput., 26 (2005), pp. 2133–2159.

[6] H. bin Zubair, C. W. Oosterlee, and R. Wienands, *Multigrid for high-dimensional elliptic partial differential equations on non-equidistant grids*, SIAM J. Sci. Comput., 29 (2007), pp. 1613–1636.

[7] M. H. Carpenter, C. Vuik, P. Lucas, M. B. van Gijzen, and H. Bijl, *A General Algorithm for Reusing Krylov Subspace Information. I. Unsteady Navier-Stokes*, NASA/Technical Memorandum 2010-216190, NASA, Langley Research Center, 2010.

[8] L. M. Carvalho, S. Gratton, R. Lago, and X. Vasseur, *A Flexible Generalized Conjugate Residual Method with Inner Orthogonalization and Deflated Restarting*, Technical Report TR/PA/10/10, CERFACS, Toulouse, France, 2010.

[9] E. de Sturler, *Nested Krylov methods based on GCR*, J. Comput. Appl. Math., 67 (1996), pp. 15–41.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[10] E. de Sturler, *Truncation strategies for optimal Krylov subspace methods*, SIAM J. Numer. Anal., 36 (1999), pp. 864–889.
[11] M. Eiermann, O. G. Ernst, and O. Schneider, *Analysis of acceleration strategies for restarted minimal residual methods*, J. Comput. Appl. Math., 123 (2000), pp. 261–292.
[12] M. Eiermann, O. G. Ernst, and E. Ullmann, *Computational aspects of the stochastic finite element method*, Comput. Visual. Sci., 10 (2007), pp. 3–15.
[13] S. C. Eisenstat, H. C. Elman, and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.
[14] J. Erhel, K. Burrage, and B. Pohl, *Restarted GMRES preconditioned by deflation*, J. Comput. Appl. Math., 69 (1996), pp. 303–318.
[15] L. Giraud, S. Gratton, X. Pinel, and X. Vasseur, *Flexible GMRES with deflated restarting*, SIAM J. Sci. Comput., 32 (2010), pp. 1858–1878.
[16] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.
[17] R. B. Morgan, *Implicitly restarted GMRES and Arnoldi methods for nonsymmetric systems of equations*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1112–1135.
[18] R. B. Morgan, *GMRES with deflated restarting*, SIAM J. Sci. Comput., 24 (2002), pp. 20–37.
[19] Y. Notay and P. S. Vassilevski, *Recursive Krylov-based multigrid cycles*, Numer. Linear Algebra Appl., 15 (2008), pp. 473–487.
[20] Y. Notay, *Flexible conjugate gradients*, SIAM J. Sci. Comput., 22 (2000), pp. 1444–1460.
[21] C. C. Paige, B. N. Parlett, and H. A. van der Vorst, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Numer. Linear Algebra Appl., 2 (1995), pp. 115–134.
[22] M. L. Parks, E. de Sturler, G. Mackey, D. D. Johnson, and S. Maiti, *Recycling Krylov subspaces for sequences of linear systems*, SIAM J. Sci. Comput., 28 (2006), pp. 1651–1674.
[23] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[24] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Statist. Comput., 14 (1993), pp. 461–469.
[25] Y. Saad, *Analysis of augmented Krylov subspace methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 435–449.
[26] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.
[27] V. Simoncini and D. B. Szyld, *Flexible inner-outer Krylov subspace methods*, SIAM J. Numer. Anal., 40 (2003), pp. 2219–2239.
[28] V. Simoncini and D. B. Szyld, *Theory of inexact Krylov subspace methods and applications to scientific computing*, SIAM J. Sci. Comput., 25 (2003), pp. 454–477.
[29] V. Simoncini and D. B. Szyld, *Recent computational developments in Krylov subspace methods for linear systems*, Numer. Linear Algebra Appl., 14 (2007), pp. 1–59.
[30] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
[31] D. B. Szyld and J. A. Vogel, *FQMR: A flexible quasi-minimal residual method with inexact preconditioning*, SIAM J. Sci. Comput., 23 (2001), pp. 363–380.
[32] A. Toselli and O. Widlund, *Domain Decomposition Methods: Algorithms and Theory*, Springer Ser. Comput. Math. 34, Springer, New York, 2004.
[33] E. Ullmann, *Solution Strategies for Stochastic Finite Element Discretizations*, Ph.D. thesis, Technische Universität Bergakademie, Freiberg, Germany, 2008.
[34] H. A. van der Vorst and C. Vuik, *GMRESR: A family of nested GMRES methods*, Numer. Linear Algebra Appl., 1 (1994), pp. 369–386.

Downloaded 09/11/12 to 194.199.1.22. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

B.4. An improved two-grid preconditioner for the solution of three-dimensional Helmholtz problems in heterogeneous media

NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS
Numer. Linear Algebra Appl. 2013; **20**:663–688
Published online 22 October 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/nla.1860

An improved two-grid preconditioner for the solution of three-dimensional Helmholtz problems in heterogeneous media‡

Henri Calandra[1], Serge Gratton[2], Xavier Pinel[3] and Xavier Vasseur[4,*,†]

[1]*TOTAL, Centre Scientifique et Technique Jean Féger, avenue de Larribau F-64000 Pau, France*
[2]*INPT-IRIT, University of Toulouse and ENSEEIHT, 2 Rue Camichel, BP 7122, F-31071 Toulouse Cedex 7, France*
[3]*CERFACS, 42 Avenue Gaspard Coriolis, F-31057 Toulouse Cedex 1, France*
[4]*CERFACS and HiePACS project joint INRIA-CERFACS Laboratory, 42 Avenue Gaspard Coriolis, F-31057 Toulouse Cedex 1, France*

SUMMARY

In this paper, we address the solution of three-dimensional heterogeneous Helmholtz problems discretized with second-order finite difference methods with application to acoustic waveform inversion in geophysics. In this setting, the numerical simulation of wave propagation phenomena requires the approximate solution of possibly very large indefinite linear systems of equations. For that purpose, we propose and analyse an iterative two-grid method acting on the Helmholtz operator where the coarse grid problem is solved inaccurately. A cycle of a multigrid method applied to a complex shifted Laplacian operator is used as a preconditioner for the approximate solution of this coarse problem. A single cycle of the new method is then used as a variable preconditioner of a flexible Krylov subspace method. We analyse the properties of the resulting preconditioned operator by Fourier analysis. Numerical results demonstrate the effectiveness of the algorithm on three-dimensional applications. The proposed numerical method allows us to solve three-dimensional wave propagation problems even at high frequencies on a reasonable number of cores of a distributed memory computer. Copyright © 2012 John Wiley & Sons, Ltd.

Received 21 December 2011; Revised 31 August 2012; Accepted 23 September 2012

KEY WORDS: complex shifted Laplacian preconditioner; flexible Krylov subspace methods; Helmholtz equation; heterogeneous media; variable preconditioning

1. INTRODUCTION

The efficient simulation of wave propagation phenomena in three-dimensional heterogeneous media is of great research interest in many environmental inverse problems (e.g. monitoring of pollution in groundwater, earthquake modelling or location of hydrocarbon in fractured rocks). Such inverse problems aim at determining accurately the material properties of the subsurface by analysing the observed scattered fields after a sequence of multiple seismic shots. One of the main computational kernels of these large-scale nonlinear optimization problems is the approximate solution of a linear system issued from the discretization of a Helmholtz scalar wave equation typically written in the frequency domain. Hence, the design of efficient iterative solvers for the resulting large indefinite linear systems is of major importance. This will be the main topic of the present paper.

When the medium is homogeneous (or similarly when the wavenumber is uniform), efficient multilevel solvers have been proposed in the literature. To name a few, we mention the wave-ray multigrid method [1], which exploits the structure of the error components that standard multigrid

*Correspondence to: Xavier Vasseur, CERFACS and HiePACS project joint INRIA-CERFACS Laboratory, 42 Avenue Gaspard Coriolis, F-31057 Toulouse Cedex 1, France.

†E-mail: xavier.vasseur@cerfacs.fr

‡Paper submitted for the special issue of Numerical Linear Algebra with Applications related to the OPTPDE ESFWaves workshop held in Würzburg, Germany, on 26–28 September 2011. Revised version.

Copyright © 2012 John Wiley & Sons, Ltd.

methods fail to eliminate [2] and the FETI-H non-overlapping domain decomposition method [3], a generalization of the FETI method [4] for Helmholtz type problems, whose rate of convergence is found to be independent of the fine grid step size, the number of subdomains, and the wavenumber in many practical problems (e.g. [5, Section 11.5.2]). In this paper, we rather focus on the case of three-dimensional Helmholtz problems defined in heterogeneous media for which the design of robust iterative methods that are scalable with respect to the frequency for such indefinite problems is currently an active research topic. The literature on iterative solvers for discrete Helmholtz problems is quite rich, and we refer the reader to the recent survey papers [6, 7] for a taxonomy of advanced preconditioned iterative methods based on domain decomposition or multigrid.

In [8], Bayliss *et al*. have considered to precondition the Helmholtz operator with a different operator. A few iterations of the symmetric successive over-relaxation method were then used to approximately invert a Laplacian preconditioner. Later, this work has been generalized by Laird and Giles [9], proposing a Helmholtz preconditioner with a positive sign in front of the Helmholtz term. In [10, 11], Erlangga *et al*. have further extended this idea: a modified Helmholtz operator with a complex wavenumber (i.e. where a complex term (hereafter named complex shift) is multiplying the square of the wavenumber) was used as a preconditioner of the Helmholtz operator. This preconditioning operator is, since then, referred to as a complex shifted Laplacian operator in the literature. This idea has received a lot of attention over the last few years; see, among others [11–13]. Indeed with an appropriate choice of the imaginary part of the shift, standard multigrid methods can be applied successfully, that is the convergence of the multigrid method as a solver or as a preconditioner applied to a complex shifted Laplacian operator is mathematically found to be mesh independent at a given frequency [14]. Nevertheless, when a multigrid method applied to a shifted Laplacian operator is considered as a preconditioner for the Helmholtz operator, the convergence is found to be frequency dependent as observed in [14, 15]. This behaviour occurs independently of the way the preconditioner is inverted (approximately or exactly). A linear increase in preconditioner applications versus the frequency is usually observed on three-dimensional problems in heterogeneous media. In practice, preconditioning based on a complex shifted Laplacian operator is considered nowadays as a successful algorithm for low to medium range frequencies.

At high frequency (or equivalently at large wavenumbers), numerical results on the contrary show a steep increase in the number of outer iterations (e.g. [14] for a concrete application in seismic imaging). The analysis of the shifted Laplace preconditioned operator provided in [13] has indeed shown that the smallest eigenvalues of the preconditioned operator tend to zero as the wavenumber increases. Hence, it becomes essential to combine this preconditioner with deflation techniques to yield an efficient numerical method as analysed in [16, 17]. As far as we know, the resulting algorithms have not yet been applied to concrete large-scale applications on realistic three-dimensional heterogeneous problems. This is indeed a topic of current research most likely due to the complexity of the numerical method. Alternatives are required and a straightforward choice considered in, for example, [18, 19] is to apply a multigrid cycle (with a limited number of grids in the hierarchy) to the Helmholtz operator. In [20], Pinel has proposed a two-grid cycle acting on the Helmholtz operator where the coarse grid problem is solved only inaccurately by a preconditioned Krylov subspace method. A theoretical analysis of this inexact preconditioner has been obtained by rigorous Fourier analysis [21] and numerical experiments on both homogeneous and heterogeneous problems have confirmed the theoretical developments. The convergence of the two-grid preconditioned Krylov subspace method was experimentally found to be mesh independent but still frequency dependent. This preconditioner has been successfully applied to the solution of huge Helmholtz problems on three-dimensional problems in heterogeneous media. Indeed numerical results reported in [20, Chapter 4] have demonstrated that the solution of large Helmholtz problems with billions of unknowns in seismic was tractable with such a two-grid preconditioned Krylov subspace method. Since then, this two-grid preconditioner has been applied to the solution of acoustic forward problems with multiple sources leading to multiple right-hand side problems [22] and to the solution of linear systems issued from the high-order discretization of the acoustic Helmholtz equation [23].

The numerical method presented in [20] is found to require a reduced number of preconditioner applications, each application being however computationally expensive. Indeed, this cycle relies on an approximate solution of a coarse problem that is highly indefinite and ill-conditioned.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

Efficient algebraic one-level preconditioners to be applied on the coarse level are missing and advanced strategies should be considered to improve the convergence properties of the two-grid approach. Hence, we propose to use a multigrid method applied to a shifted Laplacian operator as a preconditioner when solving the coarse problem. A single cycle of the new resulting method will then be used as a variable preconditioner for a flexible Krylov subspace method. By combining these two approaches, we expect an increased robustness of the numerical method and simultaneously a reduction of the computational cost of the two-grid cycle.

The contribution of this paper will be twofold. First, we will derive a new two-grid preconditioner for solving Helmholtz problems in three-dimensional heterogeneous media and analyse its properties by rigorous Fourier analysis. Second, we will show the relevance of the numerical method on a challenging application in geophysics.

The paper is organized as follows. In Section 2, we introduce the acoustic Helmholtz equation written in the frequency domain and derive the discrete linear system to be solved in the forward problem. Then in Section 3, we review two different existing preconditioners on the basis of multigrid and combine them to develop the new preconditioner. In Section 4, properties of the combined preconditioner are analysed by rigorous Fourier analysis. Furthermore, we demonstrate the effectiveness of the proposed algorithm on an academic problem and on a challenging application in geophysics in Section 5. Finally, we draw some conclusions in Section 6.

Throughout this paper, we denote by $\|.\|_2$ the Euclidean norm, $I_k \in \mathbb{C}^{k\times k}$ the identity matrix of order k and $\rho(M)$ the spectral radius of a square matrix M. Given a vector $d \in \mathbb{C}^k$ with components d_i, $D = \mathrm{diag}(d)$ is the diagonal matrix $D \in \mathbb{C}^{k\times k}$ such that $D_{ii} = d_i, (1 \leqslant i \leqslant k)$.

2. THE ACOUSTIC HELMHOLTZ EQUATION IN THE FREQUENCY DOMAIN

In this section, we briefly describe the wave propagation problem associated with acoustic imaging [24] in geophysics and introduce the mathematical formulation of this problem.

2.1. Mathematical formulation

Given a three-dimensional physical domain Ω_p of parallelepiped shape, the propagation of a wavefield in a heterogeneous medium can be modelled by the Helmholtz equation written in the frequency domain [25]:

$$-\sum_{i=1}^{3}\frac{\partial^2 u}{\partial x_i^2}-\frac{(2\pi f)^2}{c^2}u=\delta(\mathbf{x}-\mathbf{s}),\quad \mathbf{x}=(x_1,x_2,x_3)\in\Omega_p. \tag{1}$$

In Equation (1), the unknown u represents the pressure wavefield in the frequency domain, c the acoustic-wave velocity in $m\ s^{-1}$, which varies with position, and f the frequency in Hertz. The source term $\delta(\mathbf{x}-\mathbf{s})$ represents a harmonic point source located at $\mathbf{s}=(s_1,s_2,s_3)\in\Omega_p$. The wavelength λ is defined as $\lambda = c/f$ and the wavenumber as $2\pi f/c$. A popular approach—the Perfectly Matched Layer (PML) formulation [26, 27]—has been used to obtain a satisfactory near boundary solution, without many artificial reflections. Artificial boundary layers are then added around the physical domain to absorb outgoing waves at any incidence angle as shown in [26]. We denote by Ω_{PML} the surrounding domain created by these artificial layers. This formulation leads to the following set of coupled partial differential equations (PDE) with homogeneous Dirichlet boundary conditions imposed on Γ, the boundary of the domain:

$$-\sum_{i=1}^{3}\frac{\partial^2 u}{\partial x_i^2}-\frac{(2\pi f)^2}{c^2}u=\delta(\mathbf{x}-\mathbf{s})\quad\text{in}\quad\Omega_p, \tag{2}$$

$$-\sum_{i=1}^{3}\frac{1}{\xi_{x_i}(x_i)}\frac{\partial}{\partial x_i}\left(\frac{1}{\xi_{x_i}(x_i)}\frac{\partial u}{\partial x_i}\right)-\frac{(2\pi f)^2}{c^2}u=0\quad\text{in}\quad\Omega_{\mathrm{PML}}\backslash\Gamma, \tag{3}$$

$$u=0\quad\text{on}\quad\Gamma, \tag{4}$$

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

where the one-dimensional ξ_{x_i} function represents the complex-valued damping function of the PML formulation in the ith direction, selected as in [28]. The set of Equations (2)–(4) defines the forward problem related to acoustic imaging in geophysics that will be considered in this paper, and we note that the proposed numerical method can be applied to other application fields, where wave propagation phenomena appear as well.

2.2. Finite difference discretization

We use a standard second-order accurate seven-point finite difference discretization of the Helmholtz problem (2)–(4) on a uniform equidistant Cartesian grid of size $n_x \times n_y \times n_z$ (see [20, Appendix A] for a complete description of the discretization). We denote later by h the corresponding mesh grid size, Ω_h the discrete computational domain and n_{PML} the number of points in each PML layer. A fixed value of $n_{\mathrm{PML}} = 10$ has been used hereafter. Because a stability condition has to be satisfied to correctly represent the wave propagation phenomena [29], we consider a standard second-order accurate discretization scheme with 10 points per wavelength. This implies that the mesh grid size h and the minimal wavelength in the computational domain must satisfy the following inequality [29]:

$$\frac{h}{\min_{(x_1,x_2,x_3)\in\Omega_h} \lambda(x_1, x_2, x_3)} \leqslant \frac{1}{10}.$$

Hereafter, we have considered the following condition to determine the step size h, given a certain frequency f and a heterogeneous velocity field c:

$$h = \frac{\min_{(x_1,x_2,x_3)\in\Omega_h} c(x_1, x_2, x_3)}{10\, f}. \tag{5}$$

The discretization of the forward problem (2)–(4) leads to the following linear system $A_h\, x_h = b_h$, where $A_h \in \mathbb{C}^{n\times n}$ is a sparse complex matrix, which is non-Hermitian and non-symmetric because of the PML formulation [20,27,30] and where $x_h, b_h \in \mathbb{C}^n$ represent the discrete frequency-domain pressure field and source, respectively. The stability condition (5) imposes to solve large systems of equations at the (usually high) frequencies of interest for the geophysicists, a task that may be too memory expensive for standard [30, 31] or advanced sparse direct methods exploiting hierarchically semi-separable structure [32, 33] on a reasonable number of cores of a parallel computer. Consequently, preconditioned Krylov subspace methods are most often considered and efficient preconditioners must be developed for such indefinite problems. Indeed, due to the indefiniteness and the ill-conditioning of the matrices A_h, these linear systems are known to be very challenging for iterative methods [7]. Efficient preconditioners must be then developed, and, in the last years, several authors have proposed various numerical methods related to this challenging topic [12, 15, 16, 18, 34–36]. We describe next, in detail, a new iterative method proposed for the solution of the forward problem related to acoustic imaging.

3. TWO-LEVEL AND MULTI-LEVEL PRECONDITIONED KRYLOV SUBSPACE METHOD

In this section, we briefly discuss two existing preconditioning multilevel strategies for the solution of wave propagation problems presented in Section 2. Then, we introduce the new two-grid preconditioner and focus on its algorithmic description.

3.1. Two-grid cycle acting on the Helmholtz operator

We first present the general framework of the two-grid preconditioner that will serve as a basis for the new method considered in this paper and introduce some notations. The fine and coarse levels denoted by h and H are associated with discrete grids Ω_h and Ω_H, respectively. Due to the application in geophysics introduced in Section 2, where structured grids are routinely used, it seems natural to consider a geometric construction of the coarse grid Ω_H. The discrete coarse grid domain Ω_H is then deduced from the discrete fine grid domain Ω_h by doubling the mesh size in each

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

direction as classically done in vertex-centred geometric multigrid [21]. In the following, we assume that A_H represents a suitable approximation of the fine grid operator A_h on Ω_H. We also introduce $I_h^H : \mathcal{G}(\Omega_h) \to \mathcal{G}(\Omega_H)$, a restriction operator, where $\mathcal{G}(\Omega_k)$ denotes the set of grid functions defined on Ω_k. Similarly, $I_H^h : \mathcal{G}(\Omega_H) \to \mathcal{G}(\Omega_h)$ will represent a given prolongation operator. More precisely, we select as a prolongation operator, trilinear interpolation and as a restriction its adjoint, which is often called the full weighting operator [21]. We refer the reader to [37, Section 2.9] for a complete description of these operators in three dimensions.

Algorithm 1 Two-grid cycle applied to $A_h z_h = v_h$. $z_h = \mathcal{T}(v_h)$.

1: Polynomial pre-smoothing: Apply ϑ cycles of GMRES(m_s) to $A_h z_h = v_h$ with ν iterations of ω_h-Jacobi as a right preconditioner to obtain the approximation z_h^ϑ.
2: Restrict the fine level residual: $v_H = I_h^H (v_h - A_h z_h^\vartheta)$.
3: Solve approximately the coarse problem $A_H z_H = v_H$ with initial approximation $z_H^0 = 0_H$: Apply ϑ_c cycles of GMRES(m_c) to $A_H z_H = v_H$ with ν_c iterations of ω_H-Jacobi as a right preconditioner to obtain the approximation z_H.
4: Perform the coarse level correction: $\widetilde{z_h} = z_h^\vartheta + I_H^h z_H$.
5: Polynomial post-smoothing: Apply ϑ cycles of GMRES(m_s) to $A_h z_h = v_h$ with initial approximation $\widetilde{z_h}$ and ν iterations of ω_h-Jacobi as a right preconditioner to obtain the final approximation z_h.

The two-grid cycle to be used as a preconditioner is sketched in Algorithm 1, where it is assumed that the initial approximation z_h^0 is equal to zero on Ω_h, denoted later by 0_h. As in [19, 38], polynomial smoothers based on GMRES [39] have been selected for both pre-smoothing and post-smoothing phases. Here, a cycle of preconditioned GMRES(m_s) on Ω_h involves m_s matrix–vector products with A_h and $m_s \nu$ iterations of damped Jacobi. In the framework of indefinite Helmholtz problems with homogeneous velocity field, solving only approximately the coarse level problem has been analysed by rigorous Fourier analysis in [20]. Theoretical developments supported by numerical experiments have notably shown that solving approximately the coarse level problem may also lead to an efficient two-grid preconditioner. We refer the reader to [20, Section 3.4] for a complete description of this analysis on three-dimensional model problems. Finally we note that the approximation at the end of the cycle z_h can be represented as $z_h = \mathcal{T}(v_h)$, where $\mathcal{T}$ is a nonlinear function due both to the use of a polynomial method based on GMRES as a smoother and to the approximate solution obtained on the coarse grid.

3.2. Multigrid cycle acting on a complex shifted Laplacian operator

A potential drawback of the two-grid cycle acting on the Helmholtz operator presented in Section 3.1 is the indefiniteness of the coarse grid problem, which prevents from deriving an efficient multilevel method as recognized in [19]. In [11,12], Erlangga *et al*. have exploited the pioneering idea to define a preconditioning operator based on a different PDE for which a truly multilevel solution is possible. In the context of this paper, the corresponding set of equations reads as follows:

$$-\sum_{i=1}^{3} \frac{\partial^2 u}{\partial x_i^2} - (1 + i\beta)\frac{(2\pi f)^2}{c^2} u = \delta(\mathbf{x} - \mathbf{s}) \quad \text{in} \quad \Omega_p, \tag{6}$$

$$-\sum_{i=1}^{3} \frac{1}{\xi_{x_i}(x_i)} \frac{\partial}{\partial x_i} \left(\frac{1}{\xi_{x_i}(x_i)} \frac{\partial u}{\partial x_i} \right) - (1 + i\beta)\frac{(2\pi f)^2}{c^2} u = 0 \quad \text{in} \quad \Omega_{PML} \backslash \Gamma, \tag{7}$$

$$u = 0 \quad \text{on} \quad \Gamma, \tag{8}$$

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

where the parameter $1 + i\beta \in \mathbb{C}$ is called the complex shift[§]. We introduce a sequence of l grids denoted by $\Omega_1, \cdots, \Omega_l$ (with Ω_l as the finest grid) and of appropriate operators $S_k^{(\beta)}$ $(k = 1, \cdots, l)$. Here, $S_k^{(\beta)}$ is simply obtained from the second-order finite difference discretization of (6)–(8) on Ω_k. $S_k^{(\beta)}$ is later called the complex shifted Laplacian operator on Ω_k. To describe the algorithm in detail, we denote by $I_k^{k-1} : \mathcal{G}(\Omega_k) \to \mathcal{G}(\Omega_{k-1})$ a restriction operator from Ω_k to Ω_{k-1}, $I_{k-1}^k : \mathcal{G}(\Omega_{k-1}) \to \mathcal{G}(\Omega_k)$ a prolongation operator from Ω_{k-1} to Ω_k and C the cycling strategy (which can be of V, F or W type). The complex shifted multigrid algorithm considered in this paper is then sketched in Algorithm 2.

Algorithm 2 Multigrid cycle (with a hierarchy of l grids) applied to $S_l^{(\beta)} y_l = w_l$. $y_l = \mathcal{M}_{l,C}(w_l)$.

1: Pre-smoothing: Apply ν_β iterations of ω_l-Jacobi to $S_l^{(\beta)} y_l = w_l$ to obtain the approximation $y_l^{\nu_\beta}$.
2: Restrict the fine level residual: $w_{l-1} = I_l^{l-1}(w_l - S_l^{(\beta)} y_l^{\nu_\beta})$.
3: Solve approximately the coarse problem $S_{l-1}^{(\beta)} y_{l-1} = w_{l-1}$ with initial approximation $y_{l-1}^0 = 0_{l-1}$: Apply recursively γ cycles of multigrid to $S_{l-1}^{(\beta)} y_{l-1} = w_{l-1}$ to obtain the approximation y_{l-1}. On the coarsest level ($l = 1$) apply ϑ_β cycles of GMRES(m_β) preconditioned by ν_β iterations of ω_1-Jacobi to $S_1^{(\beta)} y_1 = w_1$ as an approximate solver.
4: Perform the coarse level correction: $\widetilde{y_l} = y_l^{\nu_\beta} + I_{l-1}^l y_{l-1}$.
5: Post-smoothing: Apply ν_β iterations of ω_l-Jacobi to $S_l^{(\beta)} y_l = w_l$ with initial approximation $\widetilde{y_l}$ to obtain the final approximation y_l.

In Algorithm 2, the γ parameter controls the type of cycling strategy of the multigrid hierarchy, for example, [21]. Trilinear interpolation and full-weighting are used as prolongation and restriction operators, respectively. An approximate solution on the coarsest level is considered as in the two-grid approach proposed in Section 3.1. We note that the approximation at the end of the cycle y_l can be represented as $y_l = \mathcal{M}_{l,C}(w_l)$ where $\mathcal{M}_{l,C}$ is a nonlinear function because a Krylov subspace method (namely preconditioned GMRES(m_β)) is used as an approximate solver on the coarsest grid Ω_1.

The multigrid cycle of Algorithm 2 is based on a Jacobi smoother as promoted in [12] and slightly differs from the original algorithm proposed in [12]. Indeed Erlangga *et al.* in [12] have used the matrix-dependent interpolation operator of [40], a Galerkin coarse grid approximation to deduce the discrete coarse operators and an exact solution on the coarsest grid. For three-dimensional applications, Erlangga [6] and Riyanti *et al.* [14] have proposed a multigrid method with a two-dimensional semi-coarsening strategy combined with line-wise damped Jacobi smoothing in the third direction. A cycle of multigrid acting on this complex shifted Laplacian operator is then considered as a preconditioner for the Helmholtz operator, and the theoretical properties of this preconditioner have been investigated in [13]. Since its introduction, this preconditioning technique based on a different PDE has been extensively used, see, for example , [14, 15, 34, 36] for applications in three dimensions.

3.3. Combined cycle

One of the main difficulties related to the two-grid preconditioner presented in Section 3.1 is that the coarse linear system is strongly indefinite at large wavenumbers because of the stability condition (5). Consequently, even a loose approximate solution is found to be computationally expensive to obtain with standard preconditioned Krylov subspace solvers. To circumvent this difficulty,

[§]In [11], the authors have introduced the complex shifted Laplacian with a negative imaginary part for the shift in the case of first-order or second-order radiation boundary conditions. Due to the PML formulation considered in this paper, we have used a shift with positive imaginary part to derive an efficient preconditioner as explained in [20, Section 3.3.2].

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

we introduce a multigrid cycle acting on a complex shifted Laplacian operator as a preconditioner for the coarse grid system $A_H z_H = v_H$ defined on Ω_H. The complex shifted Laplacian operator is simply obtained by direct coarse grid discretization of Equations (6)–(8) on Ω_H. The new cycle can be seen as a combination of two cycles defined on two different hierarchies. First, a two-grid cycle using Ω_h and Ω_H only as fine and coarse levels, respectively, is applied to the Helmholtz operator. Second, a sequence of grids Ω_k ($k = 1, \cdots, l$) with the finest grid Ω_l defined as $\Omega_l := \Omega_H$ is introduced. On this second hierarchy, a multigrid cycle applied to a complex shifted Laplacian operator $S_H^{(\beta)} := S_l^{(\beta)}$ is then used as a preconditioner when solving the coarse level system $A_H z_H = v_H$ of the two-grid cycle. The new combined cycle is sketched in Algorithm 3.

Algorithm 3 Combined cycle applied to $A_h z_h = v_h$. $z_h = \mathcal{T}_{l,C}(v_h)$.

1: Polynomial pre-smoothing: Apply ϑ cycles of GMRES(m_s) to $A_h z_h = v_h$ with ν iterations of ω_h-Jacobi as a right preconditioner to obtain the approximation z_h^{ϑ}.
2: Restrict the fine level residual: $v_H = I_h^H (v_h - A_h z_h^{\vartheta})$.
3: Solve approximately the coarse problem $A_H z_H = v_H$ with initial approximation $z_H^0 = 0_H$: Apply ϑ_c cycles of FGMRES(m_c) to $A_H z_H = v_H$ preconditioned by a cycle of multigrid applied to $S_l^{(\beta)} y_l = w_l$ on $\Omega_l \equiv \Omega_H$ yielding $y_l = \mathcal{M}_{l,C}(w_l)$ to obtain the approximation z_H.
4: Perform the coarse level correction: $\widetilde{z_h} = z_h^{\vartheta} + I_H^h z_H$.
5: Polynomial post-smoothing: Apply ϑ cycles of GMRES(m_s) to $A_h z_h = v_h$ with initial approximation $\widetilde{z_h}$ and ν iterations of ω_h-Jacobi as a right preconditioner to obtain the final approximation z_h.

The notation $\mathcal{T}_{l,C}$ uses subscripts related to the cycle applied to the shifted Laplacian operator (i.e. number of grids l of the second hierarchy and cycling strategy C (which can be of V, F or W type), respectively). The combined cycle then involves discretization of operators on $l + 1$ grids in total. Hence, later in the numerical experiments, we will compare $\mathcal{T}_{l,C}$ with $\mathcal{M}_{l+1,C}$. Figure 1 shows a possible configuration with a three-grid cycle applied to the shifted Laplacian operator. The combined cycle is related to the recursively defined K-cycle introduced in [41]. Nevertheless, we note that the combined cycle relies on a preconditioning operator on the coarse level that is different from the original operator. The approximation at the end of the cycle z_h can be represented as $z_h = \mathcal{T}_{l,C}(v_h)$, where $\mathcal{T}_{l,C}$ is a nonlinear function obtained as a combination of functions introduced in Sections 3.1 and 3.2, respectively. Consequently, this cycle leads to a variable nonlinear preconditioner, which must be combined with an outer *flexible* Krylov subspace method [42, 43]

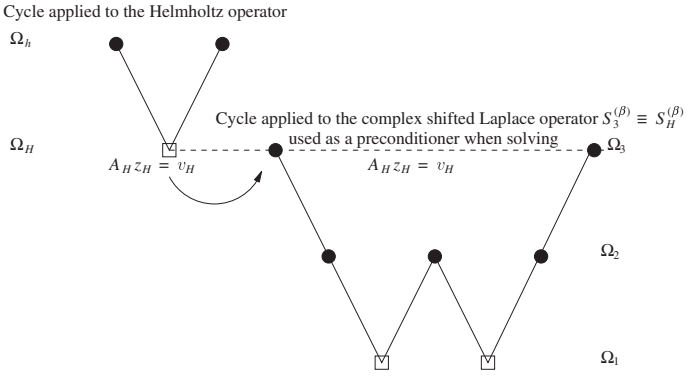


Figure 1. Combined cycle applied to $A_h z_h = v_h$ sketched in Algorithm 3. Case of $\mathcal{T}_{3,F}$. The two-grid cycle is applied to the Helmholtz operator (left part), whereas the three-grid cycle, to be used as a preconditioner when solving the coarse grid problem $A_H z_H = v_H$, is shown on the right part. This second multigrid cycle acts on the shifted Laplacian operator with β as a shift parameter.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

and [44, Chapter 10]. We have selected an outer Krylov subspace method of minimum residual type, namely flexible GMRES (FGMRES(m)) [45]. This choice allows us to characterize effectively the quality of the preconditioner even on realistic problems at a cheap cost as discussed later in Section 5.3.

4. FOURIER ANALYSIS OF MULTIGRID PRECONDITIONERS

In this section, we provide a two-grid rigorous Fourier analysis to select appropriate relaxation parameters in the smoother and to understand the convergence properties of the two-grid methods used as a preconditioner introduced in Section 3. For this analysis only, we consider a two-grid method based on a Jacobi smoother, standard coarsening, full-weighting, trilinear interpolation and exact solution on the coarse grid, applied to a model problem of Helmholtz type. We refer the reader to [21,46] for the theoretical foundations of rigorous Fourier analysis.

4.1. Rigorous Fourier analysis

Notation. Throughout Section 4, we consider the complex shifted Laplace equation with a uniform wavenumber given by $k = 2\pi f/c$ on the unit cube $\Omega = [0,1]^3$ and homogeneous Dirichlet boundary conditions on the boundary of the domain:

$$-\Delta u - \kappa_\beta^2 u = g \quad \text{in} \quad \Omega, \tag{9}$$

$$u = 0 \quad \text{on} \quad \partial\Omega, \tag{10}$$

with κ_β defined as $\kappa_\beta^2 = (1+\beta i)k^2$, where β denotes a real parameter lying in $[0,1]$. A classical tool in multigrid theory to deduce some information about the two-grid convergence rate is based on a rigorous Fourier analysis (RFA) [37, Section 3.3.4]. To perform this analysis, we introduce some additional notations. First, we discretize the model problem (9) and (10) on a uniform mesh of step size $\varkappa = 1/n_\varkappa$. We denote by $L_\varkappa^{(\beta)}$ the corresponding discrete operator on the considered fine grid $\Omega_\varkappa = G_\varkappa \cap [0,1]^3$ where $G_\varkappa$ is the infinite grid and by $D_\varkappa^{(\beta)}$ the matrix corresponding to the diagonal part of $L_\varkappa^{(\beta)}$. The discrete eigenfunctions of $L_\varkappa^{(\beta)}$:

$$\varphi_\varkappa^{l_1,l_2,l_3}(x,y,z) = \sin(l_1\pi x)\,\sin(l_2\pi y)\,\sin(l_3\pi z) \text{ with } l_1,l_2,l_3 = 1,\ldots,n_\varkappa{-}1 \text{ and } (x,y,z)\in\Omega_\varkappa,$$

generate the space of all fine grid functions, $F(\Omega_\varkappa)$, and are orthogonal with respect to the discrete inner product on $\Omega_\varkappa$:

$$(v_\varkappa, w_\varkappa) := \varkappa^3 \sum_{(x,y,z)\in\Omega_\varkappa} v_\varkappa(x,y,z) w_\varkappa(x,y,z) \quad \text{with } v_\varkappa, w_\varkappa \in F(\Omega_\varkappa).$$

The space of all fine grid real-valued functions $F(\Omega_\varkappa)$ can be divided into a direct sum of (at most) eight-dimensional subspaces—called the $2\varkappa$-harmonics [37, Equation (3.4.1)]—

$$E_\varkappa^{l_1,l_2,l_3} = \operatorname{span}\Big[\varphi_\varkappa^{l_1,l_2,l_3}, -\varphi_\varkappa^{n_\varkappa-l_1,n_\varkappa-l_2,n_\varkappa-l_3}, -\varphi_\varkappa^{n_\varkappa-l_1,l_2,l_3}, \varphi_\varkappa^{l_1,n_\varkappa-l_2,n_\varkappa-l_3},$$
$$-\varphi_\varkappa^{l_1,n_\varkappa-l_2,l_3}, \varphi_\varkappa^{n_\varkappa-l_1,l_2,n_\varkappa-l_3}, -\varphi_\varkappa^{l_1,l_2,n_\varkappa-l_3}, \varphi_\varkappa^{n_\varkappa-l_1,n_\varkappa-l_2,l_3}\Big],$$
$$\text{for } l_1,l_2,l_3 = 1,\cdots,n_\varkappa/2.$$

The dimension of $E_\varkappa^{l_1,l_2,l_3}$, denoted by $\eta_\varkappa^{l_1,l_2,l_3}$, is eight, four, two and one if zero, one, two or three of the indices l_1,l_2,l_3 is equal to $n_\varkappa/2$, respectively. Similarly as on the fine grid $\Omega_\varkappa$, we introduce the discrete eigenfunctions of the coarse grid operator $L_{2\varkappa}^{(\beta)}$ on the space of all coarse grid functions $F(\Omega_{2\varkappa})$ with $\Omega_{2\varkappa} = G_{2\varkappa}\cap[0,1]^3$:

$$\varphi_{2\varkappa}^{l_1,l_2,l_3}(x,y,z) = \sin(l_1\pi x)\,\sin(l_2\pi y)\,\sin(l_3\pi z), \text{ with } l_1,l_2,l_3 = 1,\ldots,\frac{n_\varkappa}{2}{-}1 \text{ and } (x,y,z)\in\Omega_{2\varkappa}.$$

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

$E_{2\varkappa}^{l_1,l_2,l_3}$ is then defined as $\mathrm{span}\left[\varphi_{2\varkappa}^{l_1,l_2,l_3}\right]$ because the eigenfunctions of $L_{2\varkappa}$ coincide up to their sign on $\Omega_{2\varkappa}$ for $l_1, l_2, l_3 = 1, \cdots, n_\varkappa/2$ [37]. We denote later by ℓ the multi-index $\ell = (l_1, l_2, l_3)$, by $\mathcal{L}_\varkappa = \{\ell \,|1 \leqslant \max(l_1, l_2, l_3) < n_\varkappa/2\}$ and by $\mathcal{H}_\varkappa = \{\ell \,|n_\varkappa/2 \leqslant \max(l_1, l_2, l_3) < n_\varkappa\}$ the sets of multi-indices corresponding to the low-frequency and high-frequency harmonics, respectively. We also define the set $\mathcal{L}_\varkappa^= = \{\ell \,|1 \leqslant \max(l_1, l_2, l_3) \leqslant n_\varkappa/2\}$. Later in this section, the Fourier representation of a given discrete operator $M_\varkappa$ is denoted by $\widehat{M}_\varkappa$ and the restriction of $\widehat{M}_\varkappa$ to $E_\varkappa^\ell$ with $\ell \in \mathcal{L}_\varkappa$ is noted $\widehat{M}_\varkappa(\ell) = \widehat{M}_{\varkappa|E_\varkappa^\ell}$ in short. The Fourier representation of the discrete Helmholtz operator $L_\varkappa^{(\beta)}$ and the Jacobi iteration matrix $J_\varkappa^{(\beta)}$ are denoted $\widehat{L}_\varkappa^{(\beta)}$ and $\widehat{J}_\varkappa^{(\beta)}$, respectively. To write the Fourier representation of these operators in a compact form, we also introduce the ξ_i parameters such that $\xi_i = \sin^2\left(\frac{l_i \pi \varkappa}{2}\right)$ for $i = 1, 2, 3$. Finally, we denote by $\varkappa = h$ the finest mesh grid size considered, n_h the corresponding number of points per direction and $k_\varkappa$ the wavenumber on the grid with mesh size $\varkappa$.

4.2. Smoothing analysis

The multigrid method acting on a complex shifted Laplacian operator presented in Algorithm 2 is based on a Jacobi smoother as used in [12] in two dimensions. Indeed in [12], it has been numerically shown that this method enjoys good smoothing properties on all the grids of the hierarchy when the relaxation parameters $\omega_\varkappa$ are well chosen. In Proposition 1, we give the Fourier representation of the Jacobi iteration matrix $J_\varkappa^{(\beta)}$ applied to the complex shifted Laplacian matrix $L_\varkappa^{(\beta)}$. Then, we derive related smoothing factors and by numerical experiments we deduce appropriate damping parameters to obtain good smoothing properties in three dimensions.

Proposition 1

The harmonic spaces $E_\varkappa^\ell$ for $\ell \in \mathcal{L}_\varkappa^=$ are invariant under the Jacobi iteration matrix $J_\varkappa^{(\beta)} = I_\varkappa - \omega_\varkappa \left(D_\varkappa^{(\beta)}\right)^{-1} L_\varkappa^{(\beta)}$ $\left(J_\varkappa^{(\beta)} : E_\varkappa^\ell \longrightarrow E_\varkappa^\ell, \text{ for } \ell \in \mathcal{L}_\varkappa^=\right)$. The operator $J_\varkappa^{(\beta)}$ is orthogonally equivalent to a block diagonal matrix of (at most) 8×8 blocks defined as

$$\widehat{J}_\varkappa^{(\beta)}(\ell) = I_{\eta_\varkappa^\ell} - \left(\frac{\omega_\varkappa \varkappa^2}{6 - (\kappa_\beta \varkappa)^2}\right) \widehat{L}_\varkappa^{(\beta)}(\ell), \quad \ell \in \mathcal{L}_\varkappa^=, \tag{11}$$

where $\widehat{L}_\varkappa^{(\beta)}$ denotes the representation of the complex shifted Laplacian operator $L_\varkappa^{(\beta)}$ with respect to the space $E_\varkappa^\ell$ and $\eta_\varkappa^\ell$ the dimension of $E_\varkappa^\ell$, respectively. With notation introduced in Section 4.1, if $\ell \in \mathcal{L}_\varkappa$, the representation of $\widehat{L}_\varkappa^{(\beta)}$ with respect to $E_\varkappa^\ell$ is a diagonal matrix defined as

$$\widehat{L}_\varkappa^{(\beta)}(\ell) = \mathrm{diag}\left(\frac{4}{\varkappa^2}\begin{pmatrix} (\xi_1+\xi_2+\xi_3) \\ (3-\xi_1-\xi_2-\xi_3) \\ (1-\xi_1+\xi_2+\xi_3) \\ (2+\xi_1-\xi_2-\xi_3) \\ (1+\xi_1-\xi_2+\xi_3) \\ (2-\xi_1+\xi_2-\xi_3) \\ (1+\xi_1+\xi_2-\xi_3) \\ (2-\xi_1-\xi_2+\xi_3) \end{pmatrix} \begin{matrix} -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \end{matrix}\right), \quad \ell \in \mathcal{L}_\varkappa. \tag{12}$$

If one of the indices of ℓ equals $n_\varkappa/2$, $\widehat{L}_\varkappa^{(\beta)}(\ell)$ degenerates to a diagonal matrix of dimension $\eta_\varkappa^\ell$. Its entries then correspond to the first $\eta_\varkappa^\ell$ entries of the matrix given on the right-hand side of relation (12).

Proof

Obviously, because the eigenfunctions spanning $E_\varkappa^\ell$ are eigenfunctions of $L_\varkappa^{(\beta)}$, the harmonic spaces $E_\varkappa^\ell$ ($\ell \in \mathcal{L}_\varkappa$) are invariant under $L_\varkappa^{(\beta)}$, and hence invariant under $J_\varkappa^{(\beta)}$. The representation of $L_\varkappa^{(\beta)}$

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

with respect to the harmonic space $E_\varkappa^\ell$ is obtained by writing the eigenvalues of the basis functions of $E_\varkappa^\ell$ in terms of ξ_i, a straightforward calculation that only involves trigonometric identities. □

The representation of the Jacobi iteration matrix in the Fourier space obtained in Proposition 1 allows us to easily investigate its smoothing properties, that is to compute the smoothing factor μ versus various parameters (β, mesh grid size $\varkappa$, wavenumber $k_\varkappa$ and relaxation parameter $\omega_\varkappa$, respectively). With ν denoting the number of relaxation sweeps, the smoothing factor $\mu(\beta, \varkappa, k_\varkappa, \omega_\varkappa)$ is defined as follows [47]:

$$\mu(\beta, \varkappa, k_\varkappa, \omega_\varkappa) = \max_{\ell \in \mathcal{L}_\varkappa^{\equiv}} \left| \left(\rho \left(\widehat{Q}_\varkappa(\ell) \left(\widehat{J}_\varkappa^{(\beta)}(\ell) \right)^\nu \right)^{1/\nu} \right| , \tag{13}$$

where $\widehat{Q}_\varkappa$ is the matrix representation of a projection operator that annihilates the low-frequency error components and leaves the high-frequency components unchanged [21], for example, $\widehat{Q}_\varkappa(\ell) = \mathrm{diag}((0,1,1,1,1,1,1,1)^T)$ for $\ell \in \mathcal{L}_\varkappa$. In addition, if we assume that $\kappa_\beta \varkappa$ (or similarly $k_\varkappa \varkappa$) is a given constant (which is the case in practice due to the stability condition to be satisfied), it is then possible to deduce the supremum $\mu^\star(\beta, \varkappa, k_\varkappa, \omega_\varkappa)$ of the smoothing factor over $\varkappa$ as

$$\mu^\star(\beta, \varkappa, k_\varkappa, , \omega_\varkappa) = \max \left\{ |1 - \omega_\varkappa \frac{2 - \kappa_\beta^2 \varkappa^2}{6 - \kappa_\beta^2 \varkappa^2}|, |1 - \omega_\varkappa \frac{12 - \kappa_\beta^2 \varkappa^2}{6 - \kappa_\beta^2 \varkappa^2}| \right\} , \tag{14}$$

or similarly:

$$\mu^\star(\beta, \varkappa, k_\varkappa, \omega_\varkappa) = \max \left\{ |1 - \omega_\varkappa + \frac{4\omega_\varkappa}{6 - (1 + i\beta) k_\varkappa^2 \varkappa^2}|, |1 - \omega_\varkappa - \frac{6\omega_\varkappa}{6 - (1 + i\beta) k_\varkappa^2 \varkappa^2}| \right\} . \tag{15}$$

For a fixed value of $k_\varkappa \varkappa$, this formula can then give guidance in choosing the optimal relaxation parameters and in understanding how the optimal value of the relaxation parameter $\omega_\varkappa^\star$ depends on $k_\varkappa \varkappa$ and on β, respectively. Indeed, a simple calculation gives the real-valued optimal relaxation parameter as

$$\omega_\varkappa^\star = 1 - \frac{1}{7 - k_\varkappa^2 \varkappa^2} .$$

We notice that the optimal value of the relaxation parameter does not depend on the shift parameter β, and note that we recover the optimal relaxation parameter and the supremum of the smoothing factor of the Jacobi method for the Poisson equation in three dimensions when $k_\varkappa$ is set to zero [37, Section 2.9.2].

Fourier results. We select two relaxation sweeps ($\nu = 2$) in the Jacobi method and compute the smoothing factor $\mu(\beta, \varkappa, k_\varkappa, \omega_\varkappa)$ for different values of the shift parameter β, $\omega_\varkappa$ on four consecutive grids in the multigrid hierarchy ($\varkappa = h, \varkappa = 2h, \varkappa = 4h, \varkappa = 8h$)(Figure 2). The selected wavenumbers satisfy the relation[¶] $k_\varkappa = \frac{n_h}{n_\varkappa} \frac{\pi}{5h}$ $\left(\text{or similarly } k_\varkappa = \frac{\varkappa}{h} \frac{\pi}{5h}\right)$, and we consider the case of $n_h = 512$ on the finest grid.

From Figure 2, we observe a similar behaviour as was obtained in the two-dimensional case in [12, 19]. Smoothing difficulties do occur neither on the fine grid nor on the coarsest grid of the multigrid hierarchy but on intermediate grids only. Indeed, when $\varkappa = 4h$ (bottom left part of Figure 2), smoothing factors less than one cannot be obtained unless using a complex shifted Laplace operator with $\beta \geq 0.4$. Consequently—and in agreement with the discussion provided in [12] in the two-dimensional case—we have decided to fix the shift parameter to $\beta = 0.5$. According to Figure 2, this choice leads us to consider the following relaxation parameters: $\omega_h = 0.8$, $\omega_{2h} = 0.8$, $\omega_{4h} = 0.2$, $\omega_{8h} = 1$ or in short

$$(\omega_h, \omega_{2h}, \omega_{4h}, \omega_{8h}) = (0.8, 0.8, 0.2, 1). \tag{16}$$

[¶]This corresponds to the stability condition (5) on the finest grid and to practical situations of interest on the other coarse grids of the hierarchy.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

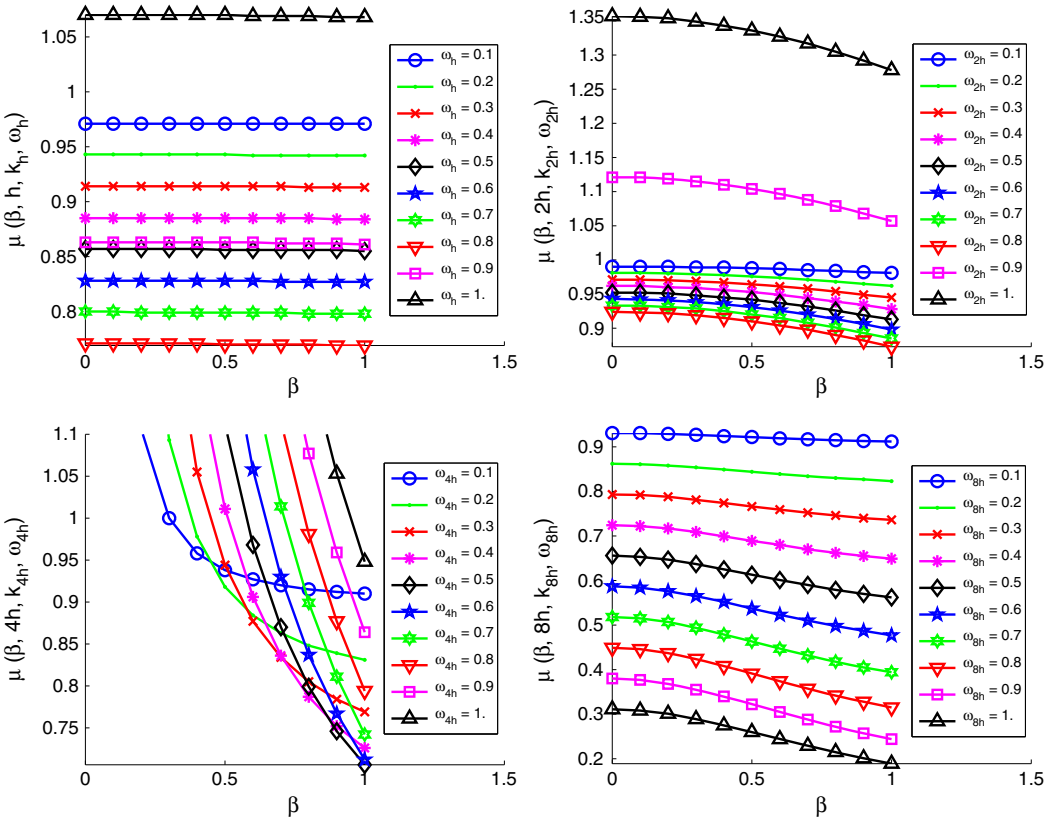


Figure 2. Smoothing factors $\mu(\beta, \varkappa, k_\varkappa, \omega_\varkappa)$ of the Jacobi method (Equation (13)) versus β and $\omega_\varkappa$ considering two relaxation sweeps ($\nu = 2$) on four different grids ($\varkappa = h$, $\varkappa = 2h$, $\varkappa = 4h$, $\varkappa = 8h$) on the model problem (9) and (10). Case of $\varkappa = h$ (top left), $\varkappa = 2h$ (top right), $\varkappa = 4h$ (bottom left) and $\varkappa = 8h$ (bottom right). The wavenumber $k_\varkappa$ is defined as $k_\varkappa = \dfrac{n_h}{n_\varkappa}\dfrac{\pi}{5h}$ with $n_h = 512$.

Table I. Optimal smoothing factors $\mu^\star(\beta, \varkappa, k_\varkappa, \omega_\varkappa)$ and optimal relaxation parameters $\omega_\varkappa^\star$ versus $\varkappa$ on the model problem (9) and (10) for two values of the shift parameter β. The wavenumber $k_\varkappa$ is defined as $k_\varkappa = \dfrac{n_h}{n_\varkappa}\dfrac{\pi}{5h}$ with h designing the stepsize of the finest grid ($n_h = 512$).

	$\beta = 0$		$\beta = 0.5$	
$\varkappa$	$\mu^\star(\beta, k_\varkappa \varkappa, \omega_\varkappa^\star)$	$\omega_\varkappa^\star$	$\mu^\star(\beta, k_\varkappa \varkappa, \omega_\varkappa^\star)$	$\omega_\varkappa^\star$
h	0.757	0.848	0.756	0.848
$2h$	0.922	0.815	0.908	0.815
$4h$	> 1		0.918	0.193
$8h$	0.274	1.055	0.231	1.055

These relaxation parameters will be selected in Section 5, and we note that they are close to the optimal values based on (15) given in Table I.

Finally, it has been shown that reasonably good smoothing factors for the Jacobi smoother can be obtained on all the grid hierarchy for the complex shifted Laplacian operator in three dimensions; see also [48], where a local Fourier analysis of the damped Jacobi method is performed on the complex shifted Laplacian in one and two dimensions. With the selected relaxation parameters, we now investigate the spectrum of preconditioned Helmholtz matrices.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

4.3. Fourier analysis of preconditioned Helmholtz operator

As shown in [47], the rigorous Fourier analysis can also provide the spectrum of a two-grid preconditioned operator inexpensively. This feature is notably quite helpful when analysing the convergence of a given preconditioned Krylov subspace method, here restarted GMRES. Next, we will perform this analysis not only on the fine level ($\varkappa = h$) to characterize the quality of the two-grid preconditioners but also on the second level ($\varkappa = 2h$) where preconditioners proposed in Algorithms 2 and 3 will be investigated. We first briefly describe how to deduce the representation of these preconditioned operators in the Fourier space.

4.3.1. Iteration matrix of a two-grid cycle.

Assumptions on the components of the cycle. In this paragraph, we assume that both the fine grid operator and the smoother leave the spaces $E_\varkappa^\ell$ invariant for $\ell \in \mathcal{L}_\varkappa^=$. As shown in Proposition 1, $L_\varkappa^{(\beta)}$ and the corresponding Jacobi iteration matrix $J_\varkappa^{(\beta)}$ do satisfy this invariance property. Furthermore, we assume that the transfer operators $I_\varkappa^{2\varkappa}$, $I_{2\varkappa}^{\varkappa}$ satisfy the following relations:

$$I_\varkappa^{2\varkappa} : E_\varkappa^\ell \to \operatorname{span}\left[\varphi_{2\varkappa}^\ell\right], I_{2\varkappa}^{\varkappa} : \operatorname{span}\left[\varphi_{2\varkappa}^\ell\right] \to E_\varkappa^\ell, \text{ for } \ell \in \mathcal{L}_\varkappa. \tag{17}$$

and that the coarse discretization operator leaves the subspace $\operatorname{span}[\varphi_{2\varkappa}^\ell]$ invariant for $\ell \in \mathcal{L}_\varkappa$. We note that the discrete coarse Helmholtz matrix $L_{2\varkappa}^{(\beta)}$ satisfies this last property and that the trilinear interpolation and its adjoint also satisfy relation (17) [37].

Proposition 2
If the previous assumptions are satisfied, the iteration matrix of the two-grid cycle ($M_\varkappa^{(\beta)} : E_\varkappa^\ell \longrightarrow E_\varkappa^\ell$, for $\ell \in \mathcal{L}_\varkappa^=$) leaves the spaces of $2\varkappa$-harmonics $E_\varkappa^\ell$ with an arbitrary $\ell \in \mathcal{L}_\varkappa^=$ invariant. The Fourier representation of the two-grid iteration matrix $M_\varkappa^{(\beta)}$ is as a block-diagonal matrix of (at most) 8×8 blocks defined as follows:

$$\widehat{M}_\varkappa^{(\beta)}(\ell) = \left(\widehat{J}_\varkappa^{(\beta)}(\ell)\right)^\nu \widehat{K}_{\varkappa,2\varkappa}^{(\beta)}(\ell) \left(\widehat{J}_\varkappa^{(\beta)}(\ell)\right)^\nu \quad \text{for} \quad \ell \in \mathcal{L}_\varkappa^=, \tag{18}$$

with $\widehat{K}_{\varkappa,2\varkappa}^{(\beta)}(\ell) = I_8 - [c\, d^T]/\Lambda_{2\varkappa}^{(\beta)}$ if $\ell \in \mathcal{L}_\varkappa$, where $\Lambda_{2\varkappa}^{(\beta)} = \frac{4}{\varkappa^2}((1-\xi_1)\xi_1 + (1-\xi_2)\xi_2 + (1-\xi_3)\xi_3) - \kappa_\beta^2$ and $c \in \mathbb{R}^8$, $d \in \mathbb{C}^8$, are defined as follows:

$$\begin{cases} c_1 = (1-\xi_1)(1-\xi_2)(1-\xi_3),\ c_2 = \xi_1\xi_2\xi_3,\ c_3 = \xi_1(1-\xi_2)(1-\xi_3),\ c_4 = (1-\xi_1)\xi_2\xi_3, \\ c_5 = (1-\xi_1)\xi_2(1-\xi_3),\ c_6 = \xi_1(1-\xi_2)\xi_3,\ c_7 = (1-\xi_1)(1-\xi_2)\xi_3,\ c_8 = \xi_1\xi_2(1-\xi_3), \\ d = \widehat{L}_\varkappa^{(\beta)}(\ell)\, c, \text{ where } \widehat{L}_\varkappa^{(\beta)}(\ell) \text{ is defined in Equation (12).} \end{cases}$$

If one of the indices of ℓ is equal to $n_\varkappa/2$, $\widehat{K}_{\varkappa,2\varkappa}^{(\beta)}(\ell)$ is reduced to the identity matrix of dimension $\eta_\varkappa^\ell$.

Proof
Under the assumptions given earlier, it is straightforward to prove that the iteration matrix of the two-grid cycle leaves $E_\varkappa^\ell$ for $\ell \in \mathcal{L}_\varkappa^=$ invariant. We obtain formula (18) by just combining the Fourier representation of each of its components. The complete details of these trigonometric calculations can be found in [20, Section 3.3.1]. □

4.3.2. Fourier representation of preconditioned Helmholtz operator.

In this paragraph, we consider the solution of the following linear system $L_\varkappa^{(\sigma_L)} y_\varkappa = w_\varkappa$ with a given Krylov subspace method. The corresponding matrix $L_\varkappa^{(\sigma_L)}$ is a possibly complex shifted Laplacian matrix with $\kappa_{\sigma_L}^2 = (1 + i\sigma_L)k_\varkappa^2 \in \mathbb{C}$, $k_\varkappa = \dfrac{n_h}{n_\varkappa}\dfrac{\pi}{5h}$, where $\varkappa$ is the mesh grid size and σ_L denotes a shift parameter lying in [0, 1] . The preconditioning matrix can be a two-grid iteration matrix $M_\varkappa^{(\sigma_p)}$ or a

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

Jacobi iteration matrix $J_\varkappa^{(\sigma_p)}$, both applied to a possibly complex shifted Laplacian operator $L_\varkappa^{(\sigma_p)}$ with $\kappa_{\sigma_p}^2 = (1 + i\sigma_p)k_\varkappa^2$, where σ_p denotes a shift parameter lying in $[0, 1]$. Each preconditioning step requires an approximate solution of the linear system $L_\varkappa^{(\sigma_p)} z_\varkappa = v_\varkappa$. If one cycle of a geometric two-grid method is used to approximate the inverse of $L_\varkappa^{(\sigma_p)}$, we denote by $\mho_\varkappa^{-1}(\sigma_p)$ this approximation. Similarly, if ν relaxation sweeps of a Jacobi method are used to approximate the inverse of $L_\varkappa^{(\sigma_p)}$, we denote by $\Upsilon_\varkappa^{-1}(\sigma_p)$ this approximation. The convergence of the Krylov subspace method with right preconditioning is partly related to the spectra of the matrices $L_\varkappa^{(\sigma_L)} \mho_\varkappa^{-1}(\sigma_p)$ or $L_\varkappa^{(\sigma_L)} \Upsilon_\varkappa^{-1}(\sigma_p)$. As shown in [47], the iteration matrices of both preconditioning phases correspond to

$$M_\varkappa^{(\sigma_p)} = \left(I_\varkappa - \mho_\varkappa^{-1}(\sigma_p)\, L_\varkappa^{(\sigma_p)}\right) \quad \text{or} \quad \mho_\varkappa^{-1}(\sigma_p)\, L_\varkappa^{(\sigma_p)} = I_\varkappa - M_\varkappa^{(\sigma_p)}, \tag{19}$$

$$J_\varkappa^{(\sigma_p)^\nu} = \left(I_\varkappa - \Upsilon_\varkappa^{-1}(\sigma_p)\, L_\varkappa^{(\sigma_p)}\right) \quad \text{or} \quad \Upsilon_\varkappa^{-1}(\sigma_p)\, L_\varkappa^{(\sigma_p)} = I_\varkappa - J_\varkappa^{(\sigma_p)^\nu}. \tag{20}$$

From (19) and (20), the following relations can be easily deduced :

$$L_\varkappa^{(\sigma_L)} \mho_\varkappa^{-1}(\sigma_p) = L_\varkappa^{(\sigma_L)} \left(I_\varkappa - M_\varkappa^{(\sigma_p)}\right) \left(L_\varkappa^{(\sigma_p)}\right)^{-1}, \tag{21}$$

$$L_\varkappa^{(\sigma_L)} \Upsilon_\varkappa^{-1}(\sigma_p) = L_\varkappa^{(\sigma_L)} \left(I_\varkappa - J_\varkappa^{(\sigma_p)^\nu}\right) \left(L_\varkappa^{(\sigma_p)}\right)^{-1}. \tag{22}$$

Remark. Since all operators in Equation (21) are block diagonal in the Fourier space (see Propositions 1 and 2, respectively), the spectrum of $L_\varkappa^{(\sigma_L)} \mho_\varkappa^{-1}(\sigma_p)$ is obtained by solving eigenvalue problems of small dimension only (8 at most). This is inexpensive. We also remark that the Fourier representation of $L_\varkappa^{(\sigma_L)} \Upsilon_\varkappa^{-1}(\sigma_p)$ is a diagonal matrix (Proposition 1), its spectrum is then obtained straightforwardly.

4.3.3. Fourier results.

Fine level $\varkappa = h$—Figure 3. We first analyse the spectrum of $L_h^{(\sigma_L)} \mho_h^{-1}(\sigma_p)$ for $\sigma_L = 0$ (i.e. the Helmholtz operator) with two different preconditioners. We will consider the case of a preconditioner on the basis of a two-grid method acting either on the Helmholtz operator ($\sigma_p = 0$) or on a complex shifted Laplacian operator ($\sigma_p = 0.5$). The corresponding spectra of $L_h^{(0)} \mho_h^{-1}(0)$ and $L_h^{(0)} \mho_h^{-1}(0.5)$ are shown in Figure 3.

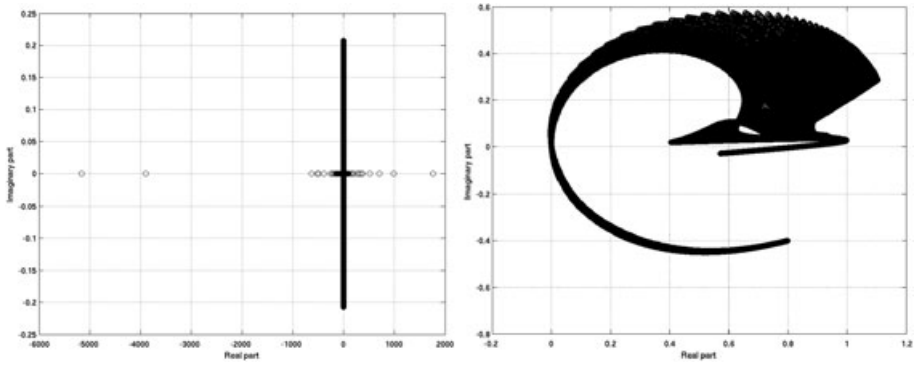


Figure 3. Spectra of $L_h^{(0)} \mho_h^{-1}(\sigma_p)$ for two different two-grid preconditioners ($\sigma_p = 0$, $\omega_h = 0.8$, $\nu = 2$) (left part) and ($\sigma_p = 0.5$, $\omega_h = 0.8$, $\nu = 2$) (right part), with $h = \dfrac{1}{256}$ for a wavenumber such as $k_h = \pi/(5h)$. Note the different scales used in both figures.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

Using the two-grid method on the Helmholtz operator leads to a spectrum with a cluster around $(1, 0)$ in the complex plane with relatively a few isolated eigenvalues with both positive and negative real parts (left part of Figure 3). When the two-grid method is applied to the complex shifted Laplacian matrix, the spectrum shown on the right part of Figure 3 is lying in the positive real part of the complex plane only with relatively few eigenvalues close to zero (less than 0.1% of the spectrum is located inside the disc of radius 0.1 centred at the origin). Moreover, it has to be noticed that the shapes of these spectra are similar to those obtained in two dimensions; see Figure 1 in [49] for the Helmholtz matrix and Figure 7 in [12] (up to a symmetry with respect to the x-axis) for the complex shifted Laplacian matrix, respectively. Both spectra relatively look in favour of the convergence of a Krylov subspace method as will be confirmed by numerical experiments on a homogeneous Helmholtz problem in Section 5.2.

Coarse level $\varkappa = 2h$—Figure 4. We now study the preconditioner properties on the coarse level in a two-grid method acting on the Helmholtz problem. More precisely, we consider two different preconditioners to solve the indefinite coarse problem approximately. First, two iterations ($\nu = 2$) of damped Jacobi with $\omega_{2h} = 0.8$ are used as a preconditioner of the coarse Helmholtz matrix (step 3 of Algorithm 1). The spectrum of $L_{2h}^{(0)}\, \Upsilon_{2h}^{-1}(0)$ is shown on the left part of Figure 4. Second, a complex shifted multigrid method is used to solve approximately the coarse Helmholtz problem (step 3 of Algorithm 3). The spectrum of the preconditioned coarse Helmholtz matrix $L_{2h}^{(0)}\, \mho_{2h}^{-1}(0.5)$ is shown on the right part of Figure 4.

If we compare the two plots related to the complex shifted multigrid preconditioner (right parts of Figures 3 and 4, respectively), we remark that both spectra have a similar curved shape. Most of the eigenvalues have a real part located between 0. and 1.2, whereas only a few outliers have a negative real part close to zero. A similar behaviour in terms of convergence is then expected on both fine and coarse levels when such a preconditioner is used. On the opposite, the Jacobi coarse preconditioner acts quite differently. No cluster appears in the spectrum shown on the left part of Figure 4 and even worse the real part of the eigenvalues is located between 0 and 2 million with a few outliers having a negative real part close to zero. This spread of eigenvalues in the spectrum may strongly penalize the convergence of GMRES on the coarse level ($\varkappa = 2h$). Consequently, according to both spectra shown in Figure 4, the preconditioner based on a cycle of multigrid applied to a complex shifted Laplacian operator seems to be a more appropriate choice to solve the coarse Helmholtz problem approximately.

Coarse level $\varkappa = 4h$—Figure 5. We conclude this analysis by studying the properties of the Jacobi preconditioner on the coarsest level ($\varkappa = 4h$) in a complex shifted multigrid cycle (step 3 of Algorithm 2). The spectrum of $L_{4h}^{(\sigma_L)}\, \Upsilon_{4h}^{-1}(\sigma_p)$ is shown in Figure 5 for $\sigma_L = \sigma_p = 0.5$ with

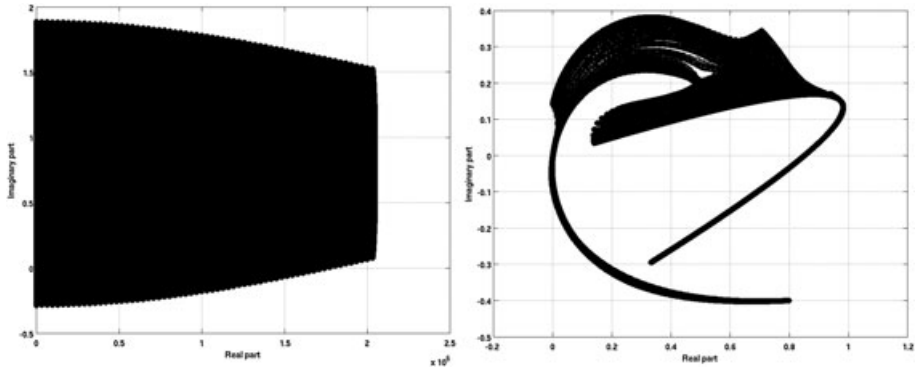


Figure 4. Spectrum of $L_{2h}^{(0)}\, \Upsilon_{2h}^{-1}(\sigma_p)$ ($\sigma_p = 0$, $\omega_r = 0.8$, $\nu = 2$) (left part) and of $L_{2h}^{(0)}\, \mho_{2h}^{-1}(\sigma_p)$ ($\sigma_p = 0.5$, $\omega_{2h} = 0.8$, $\nu = 2$) (right part), with $h = \dfrac{1}{256}$ for a wavenumber k_{2h} such that $k_{2h} = 2\pi/(5h)$. Note the different scales used in both figures.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

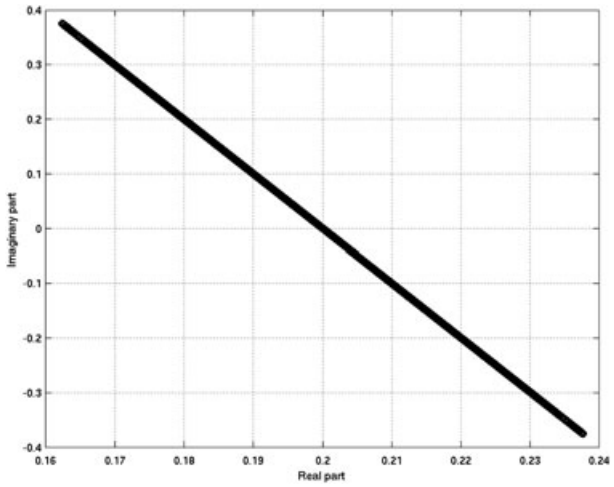


Figure 5. Spectrum of $L_{4h}^{(\sigma_L)}\,\Upsilon_{4h}^{-1}(\sigma_p)$, with $\sigma_L = \sigma_p = 0.5$, $\omega_{4h} = 0.2$, $\nu = 2$, $h = \dfrac{1}{256}$ on a 64^3 grid with $k_{4h} = 4\pi/(5h)$.

$\nu = 2$ relaxation sweeps of damped Jacobi ($\omega_{4h} = 0.2$) as a preconditioner. This spectrum looks in favour of the convergence of GMRES. Indeed the preconditioned matrix $L_{4h}^{(0.5)}\,\Upsilon_{4h}^{-1}(0.5)$ is actually a positive definite complex matrix and satisfies a sufficient condition to ensure the convergence of GMRES [50, Theorem 6.30].

4.4. Rigorous Fourier analysis for operators with variable coefficients

In this subsection, we consider only the complex shifted Laplace equation now with smoothly variable coefficients on the unit cube $\Omega = [0,1]^3$ and homogeneous Dirichlet boundary conditions on the boundary of the domain:

$$-\sum_{i=1}^{3} \frac{1}{\xi_{x_i}(x_i)} \frac{\partial}{\partial x_i}\left(\frac{1}{\xi_{x_i}(x_i)} \frac{\partial u}{\partial x_i}\right) - \kappa_\beta^2 u = g \quad \text{in} \quad \Omega, \tag{23}$$

$$u = 0 \quad \text{on} \quad \partial\Omega. \tag{24}$$

This model problem aims at representing the partial differential equation to be solved when using the PML formulation. We denote by $L_\varkappa^{(\beta)}(x)$ and $D_\varkappa^{(\beta)}(x)$ the discretized operator with variable coefficients on the considered fine grid $\Omega_\varkappa$ and its diagonal part, respectively. A direct application of rigorous Fourier analysis is not possible for PDE with variable coefficients [37,51]. The smoothing factor indeed becomes x-dependent. However, the analysis can be applied to the locally frozen operator $L_\varkappa^{(\beta)}(x_f)$ at a fixed grid point $x_f \in \Omega$ of coordinates $(x_{f_1}, x_{f_2}, x_{f_3})$: it reduces to perform the rigorous Fourier analysis on the operator with frozen coefficients. To perform such analysis, we now assume that the finite difference stencil notation of the discretized operator $L_\varkappa^{(\beta)}(x_f)$ can be written as

$$L_{\varkappa,(0)}^{(\beta)}(x_f) = \frac{1}{\varkappa^2}\begin{bmatrix} & -\chi_2 & \\ -\chi_1 & 2(\chi_1+\chi_2+\chi_3) - (\kappa_\beta \varkappa)^2 & -\chi_1 \\ & -\chi_2 & \end{bmatrix}$$

$$L_{\varkappa,(-1)}^{(\beta)}(x_f) = \frac{1}{\varkappa^2}\begin{bmatrix} & -\chi_3 & \end{bmatrix}, \quad L_{\varkappa,(1)}^{(\beta)}(x_f) = \frac{1}{\varkappa^2}\begin{bmatrix} & -\chi_3 & \end{bmatrix}$$

with $\chi_i \in \mathbb{C}$ $(i = 1,2,3)$ defined as $1/(\xi_{x_i}^2(x_{f_i}))$. First, we extend Proposition 1 to the case of PDE with variable coefficients.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

4.4.1. Smoothing factor.

Proposition 3
At a given point $x_f \in \Omega_\varkappa$ the harmonic spaces $E_\varkappa^\ell$ for $\ell \in \mathcal{L}_\varkappa^=$ are invariant under the Jacobi iteration matrix $J_\varkappa^{(\beta)}(x_f) = I_\varkappa - \omega_\varkappa (D_\varkappa^{(\beta)}(x_f))^{-1} L_\varkappa^{(\beta)}(x_f)$ ($J_\varkappa^{(\beta)}(x_f) : E_\varkappa^\ell \longrightarrow E_\varkappa^\ell$, for $\ell \in \mathcal{L}_\varkappa^=$). The operator $J_\varkappa^{(\beta)}(x_f)$ is orthogonally equivalent to a block diagonal matrix of (at most) 8×8 blocks defined as

$$\widehat{J}_\varkappa^{(\beta)}(\ell, x_f) = I_{\eta_\varkappa^\ell} - \left(\frac{\omega_\varkappa \varkappa^2}{2(\chi_1 + \chi_2 + \chi_3) - (\kappa_\beta \varkappa)^2} \right) \widehat{L}_\varkappa^{(\beta)}(\ell, x_f), \ell \in \mathcal{L}_\varkappa^=, \tag{25}$$

where $\widehat{L}_\varkappa^{(\beta)}(x_f)$ denotes the representation of the complex shifted Laplacian operator $L_\varkappa^{(\beta)}(x_f)$ with respect to the space $E_\varkappa^\ell$ at point x_f and $\eta_\varkappa^\ell$ the dimension of $E_\varkappa^\ell$, respectively. With notation introduced in Section 4.1, if $\ell \in \mathcal{L}_\varkappa$, the representation of $\widehat{L}_\varkappa^{(\beta)}(x_f)$ with respect to $E_\varkappa^\ell$ is a diagonal matrix defined as

$$\widehat{L}_\varkappa^{(\beta)}(\ell, x_f) = \operatorname{diag}\left(\frac{4}{\varkappa^2} \begin{pmatrix} (\chi_1\xi_1 + \chi_2\xi_2 + \chi_3\xi_3) \\ (\chi_1(1-\xi_1) + \chi_2(1-\xi_2) + \chi_3(1-\xi_3)) \\ (\chi_1(1-\xi_1) + \chi_2\xi_2 + \chi_3\xi_3) \\ (\chi_1\xi_1 + \chi_2(1-\xi_2) + \chi_3(1-\xi_3)) \\ (\chi_1\xi_1 + \chi_2(1-\xi_2) + \chi_3\xi_3) \\ (\chi_1(1-\xi_1) + \chi_2\xi_2 + \chi_3(1-\xi_3)) \\ (\chi_1\xi_1 + \chi_2\xi_2 + \chi_3(1-\xi_3)) \\ (\chi_1(1-\xi_1) + \chi_2(1-\xi_2) + \chi_3\xi_3) \end{pmatrix} \begin{matrix} -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \\ -\kappa_\beta^2 \end{matrix} \right). \tag{26}$$

If one of the indices of ℓ equals $n_\varkappa/2$, $\widehat{L}_\varkappa^{(\beta)}(\ell, x_f)$ degenerates to a diagonal matrix of dimension $\eta_\varkappa^\ell$. Its entries then correspond to the first $\eta_\varkappa^\ell$ entries of the matrix given on the right-hand side of relation (26).

Proof
This can be obtained by using trigonometric identities as in Proposition 1. □

The worst-case smoothing factor $\mu_{wc}(\beta, \varkappa, k_\varkappa, \omega_\varkappa)$ is then defined as

$$\begin{aligned} \mu_{wc}(\beta, \varkappa, k_\varkappa, \omega_\varkappa) &= \max_{x_f \in \Omega} \mu(\beta, \varkappa, k_\varkappa, \omega_\varkappa, x_f), \\ &= \max_{x_f \in \Omega} \max_{\ell \in \mathcal{L}_\varkappa^=} |(\rho(\widehat{Q}_\varkappa(\ell)\ (\widehat{J}_\varkappa^{(\beta)}(\ell, x_f))^\nu)^{1/\nu}|. \end{aligned}$$

As an illustration, Figure 6 shows the smoothing factors at a selected point x_f chosen in the PML layer such as $1/\chi_j = (1 + i\cos(3\pi/8))^2, (j = 1, 2, 3)$ (with a PML function selected as in [28]) for different values of β on four different grids ($\varkappa = h$, $\varkappa = 2h$, $\varkappa = 4h$, $\varkappa = 8h$, respectively) with a wavenumber defined as $k_\varkappa = \dfrac{n_h}{n_\varkappa}\dfrac{\pi}{5h}$ with $n_h = 512$. For such a choice of the χ_j coefficients, we note that smoothing factors less than one can be obtained on the intermediate coarse grid $\varkappa = 4h$ whatever β. When β is set to 0.5, Table II reveals that reasonable worst-case values of the smoothing factors can be obtained on the different grids as in Section 4.2. On the other hand, for the considered combination of $\omega_\varkappa$, $k_\varkappa$ and $\varkappa$, it is possible to obtain worst-case values of the smoothing factor greater than one when β is equal to 0; this justifies the use of a Krylov acceleration procedure as a smoother, as recommended in [19].

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

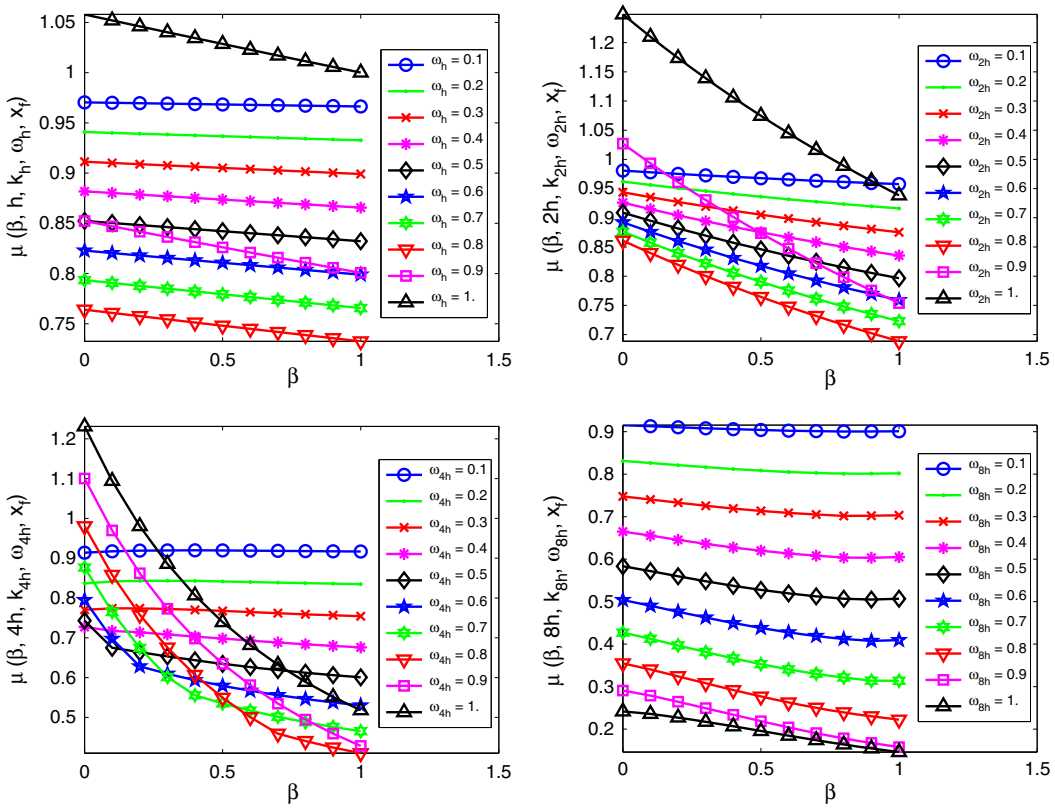


Figure 6. Smoothing factors $\mu(\beta, \varkappa, k_\varkappa, \omega_\varkappa, x_f)$ of the Jacobi method versus β and $\omega_\varkappa$ considering 2 relaxation sweeps ($\nu = 2$) on four different grids ($\varkappa = h$, $\varkappa = 2h$, $\varkappa = 4h$, $\varkappa = 8h$) on the model problem (23) and (24). Case of $\varkappa = h$ (top left), $\varkappa = 2h$ (top right), $\varkappa = 4h$ (bottom left) and $\varkappa = 8h$ (bottom right). The wavenumber $k_\varkappa$ is defined as $k_\varkappa = \dfrac{n_h}{n_\varkappa}\dfrac{\pi}{5h}$ with $n_h = 512$. Case of $1/\chi_j = (1 + i\cos(3\pi/8))^2$, $(j = 1, 2, 3)$.

Table II. Computed worst-case smoothing factors $\mu_{\mathrm{wc}}(\beta, \varkappa, k_\varkappa, \omega_\varkappa)$ for the model problem (23) and (24) for two values of β versus $\varkappa$. The wavenumber $k_\varkappa$ is defined as $k_\varkappa = \frac{n_h}{n_\varkappa}\frac{\pi}{5h}$ with $n_h = 512$.

	$\beta = 0$		$\beta = 0.5$	
$\varkappa$	$\mu_{wc}(\beta, \varkappa, k_\varkappa, \omega_\varkappa)$	$\omega_\varkappa$	$\mu_{wc}(\beta, \varkappa, k_\varkappa, \omega_\varkappa)$	$\omega_\varkappa$
h	0.791	0.8	0.770	0.8
$2h$	> 1	0.8	0.914	0.8
$4h$	> 1	0.2	0.918	0.2
$8h$	0.310	1	0.260	1

4.5. *Conclusions*

To conclude, we have selected, with the rigorous Fourier analysis, appropriate relaxation parameters in the Jacobi method that lead to acceptable smoothing factors on all the grids of a complex shifted multigrid method in three dimensions (Figure 3). As a new result, we have shown the suitability of the complex shifted multigrid preconditioner on the coarse level of a combined two-grid method (left part of Figure 4). Finally, we have also demonstrated the good preconditioning properties of a Jacobi preconditioner on the coarsest level of a complex shifted multigrid (Figure 5). Although rigorous Fourier analysis corresponds to a simplified analysis, numerical experiments detailed in Section 5 will support these conclusions.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

5. NUMERICAL EXPERIMENTS ON THREE-DIMENSIONAL PROBLEMS

We investigate the performance of the various preconditioners presented in Section 3 combined with Flexible GMRES(m) for the solution of the acoustic Helmholtz problem (2)–(4) on a homogeneous problem and on a realistic heterogeneous velocity model.

5.1. Settings

In the two-grid cycle of Algorithm 1, we consider as a smoother the case of one cycle of GMRES(2) preconditioned by two iterations of damped Jacobi ($\vartheta = 1$, $m_s = 2$ and $\nu = 2$), a restarting parameter equal to $m_c = 10$ for the preconditioned GMRES method used on the coarse level and a maximal number of coarse cycles equal to $\vartheta_c = 10$. In the complex shifted multigrid cycle of Algorithm 2, we use a shift parameter equal to $\beta = 0.5$ and two iterations of damped Jacobi as a smoother ($\nu_\beta = 2$). On the coarsest level, we consider, as an approximate solver, one cycle of GMRES(10) preconditioned by two iterations of damped Jacobi ($\vartheta_\beta = 1$, $m_\beta = 10$ and $\nu_\beta = 2$). The previous parameters have also been used in Algorithm 3, exception made for ϑ_c set to 2. Finally, the relaxation coefficients considered in the Jacobi method have been determined by rigorous Fourier analysis and are given by relation (16).

We consider a value of the restarting parameter of the outer Krylov subspace method equal to $m = 5$ as in [20, 22] (see the first remark given in Section 5.3 for further comments). The unit source is located at $(s_1, s_2, s_3) = (h\ n_{x_1}/2, h\ n_{x_2}/2, h\ (n_{PML} + 1))$ where,for example, n_{x_1} denotes the number of points in the first direction. A zero initial guess x_h^0 is chosen and the iterative method is stopped when the Euclidean norm of the residual normalized by the Euclidean norm of the right-hand side satisfies the following relation:

$$\frac{||b_h - A_h x_h||_2}{||b_h||_2} \leqslant 10^{-5}. \tag{27}$$

The numerical results have been obtained on Babel, an IBM Blue Gene/P computer located at IDRIS (each node of Babel is equipped with 4 PowerPC 450 cores at 850 Mhz) using a Fortran 90 implementation with MPI [52] in complex single precision arithmetic (see [37, Chapter 6] for the practical aspects related to the parallelization of geometric multigrid). Physical memory on a given node (four cores) of Babel is limited to 2 GB. This code was compiled by the IBM compiler suite with the best optimization options and linked with the vendor BLAS and LAPACK subroutines.

5.2. Homogeneous velocity field

We consider the case of a homogeneous velocity field in a reference domain $[0, 1]^3$ as a first benchmark problem. The step size of the Cartesian mesh of type n_h^3 is given by $h = 1/n_h$ and a uniform wavenumber k is imposed such that $kh = \pi/5$ as stated in relation (5). Consequently, large wavenumbers are obtained when the step size h is small. Table III collects the number of preconditioner applications (Prec), computational times (T) and maximal requested memory (M) for the various preconditioners investigated in Section 3: a two-grid preconditioner ($\mathcal{T}$), two four-grid complex shifted preconditioners ($\mathcal{M}_{4,V}$ and $\mathcal{M}_{4,F}$) and three variants of two-grid cycles with complex shifted two-grid cycle ($\mathcal{T}_{2,V}$) or three-grid cycles ($\mathcal{T}_{3,V}$ and $\mathcal{T}_{3,F}$) as a coarse preconditioner, respectively. Finally the number of cores (# Cores) is selected such that the dimension of the local problem on the finest grid is fixed for a given strategy in these numerical experiments.

The number of preconditioner applications (Prec) is found to grow almost linearly with the wavenumber, whatever the preconditioning strategies. This behaviour has already been pointed out in [12, 14, 15, 53] for the complex shifted preconditioner in two-dimensional and three-dimensional applications, when addressing problems of smaller dimension, however. We note that the two-grid cycles used as a preconditioner usually require a moderate number of preconditioner applications (each application being however computationally expensive). As expected, using the combined cycles ($\mathcal{T}_{2,V}$, $\mathcal{T}_{3,V}$ or $\mathcal{T}_{3,F}$) leads to a significant decrease in terms of computational times with respect to the two-grid preconditioner ($\mathcal{T}$) initially proposed in [20]: a reduction factor of at least 1.5 is obtained even at high wavenumbers. This can be considered as a noticeable improvement.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

Table III. Preconditioned flexible methods for the solution of the Helmholtz equation for the homogeneous velocity field. Case of a second-order discretization with 10 points per wavelength such that $kh = \pi/5$. Case of two-grid ($\mathcal{T}$), of complex shifted multigrid cycles ($\mathcal{M}_{4,V}$, $\mathcal{M}_{4,F}$) and of combined cycles ($\mathcal{T}_{2,V}$, $\mathcal{T}_{3,V}$ and $\mathcal{T}_{3,F}$) applied as a preconditioner of FGMRES(5).

Homogeneous velocity field							
Grid	# Cores	Prec	T (s)	M (GB)	Prec	T (s)	M (GB)
		$\mathcal{T}$			$\mathcal{T}_{2,V}$		
128^3	1	18	455	0.3	17	309	0.3
256^3	8	29	790	2.4	28	552	2.4
512^3	64	49	1354	19.2	52	1047	19.5
1024^3	512	92	2588	154.0	100	2067	155.7
2048^3	4096	228	6593	1232.0	207	4447	1245.5
		$\mathcal{M}_{4,V}$			$\mathcal{M}_{4,F}$		
128^3	1	95	251	0.3	125	372	0.3
256^3	8	180	505	2.0	180	573	2.0
512^3	64	355	1026	16.4	339	1107	16.4
1024^3	512	696	2112	130.8	635	2165	130.8
2048^3	4096	1415	4644	1046.8	1278	4634	1046.8
		$\mathcal{T}_{3,V}$			$\mathcal{T}_{3,F}$		
128^3	1	17	**250**	0.3	18	289	0.3
256^3	8	29	**463**	2.4	30	528	2.4
512^3	64	54	**877**	19.6	56	1007	19.6
1024^3	512	105	**1746**	157.1	107	1980	157.1
2048^3	4096	259	**4442**	1256.5	247	4752	1256.5

Prec, the number of preconditioner applications; T, the total computational time in seconds; M, the requested memory in GB.
The bold values represent the minimal computational times, which $\mathcal{T}_{3,V}$ strategy always delivers.
Numerical experiments were performed on an IBM BG/P computer.

Furthermore, we notice that the numbers of preconditioner applications obtained with the combined approaches are almost similar. Using a coarse preconditioner with a hierarchy of three grids such as in $\mathcal{T}_{3,V}$ or $\mathcal{T}_{3,F}$ allows us to reduce the computational times with respect to the $\mathcal{T}_{2,V}$ approach. Concerning the complex shifted preconditioners, we remark that the $\mathcal{M}_{4,V}$ strategy performs well in terms of computational times with respect to $\mathcal{M}_{4,F}$. Indeed, on this homogeneous problem, a preconditioner such as the V-cycle, with a cycling strategy visiting the coarsest level only once, seems to be a good compromise in terms of computational times. Among the six investigated preconditioning strategies, $\mathcal{T}_{3,V}$ always delivers the minimal computational times (see bold values in Table III). Compared with $\mathcal{M}_{4,V}$, $\mathcal{T}_{3,V}$ leads to a reduction in terms of computational times of about 20% (1024^3) and of 4.5% on the largest test case (2048^3). Finally, we note that the maximal requested memory (M) grows linearly with the problem size whatever the preconditioner. This is indeed the expected behaviour, because these strategies do not rely on any (local or global) factorization of sparse matrices. The complex shifted preconditioners $\mathcal{M}_{4,F}$ and $\mathcal{M}_{4,V}$ require less memory than the combined strategies $\mathcal{T}_{3,V}$ and $\mathcal{T}_{3,F}$: a factor of reduction of 20% is indeed observed. Furthermore, we point out that the numerical methods investigated in this paper on both homogeneous or heterogeneous cases are relatively cheap in terms of memory requirements, for example, an amount of only 157 GB at most is needed when solving a wave propagation problem with more than one billion of unknowns (1024^3). This feature is especially important when addressing in a near future the solution of multiple right-hand side problems arising in the related acoustic imaging inverse problem.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

5.3. EAGE/SEG salt dome

The SEG/EAGE salt dome model [54] is a velocity field containing a salt dome in a sedimentary embankment. It is defined in a parallelepiped domain of size $13.5 \times 13.5 \times 4.2$ km^3. The minimum value of the velocity is 1500 m.s^{-1}, and its maximum value is 4481 m.s^{-1}, respectively. This test case is considered challenging due to both the occurrence of a geometrically complex structure (salt dome) and the large dimensions of the computational domain.

We are mostly interested in evaluating the behaviour of the different preconditioners versus the frequency on this heterogeneous velocity field problem. We consider a set of frequencies ranging from 2.5 Hz to 40 Hz with a step size h selected such that the stability condition (5) is satisfied. We note that the largest frequency case ($f = 40$ Hz) corresponds to a linear system of approximately 15.8 billion of unknowns. In the numerical experiments, we analyse four different strategies: a two-grid preconditioner ($\mathcal{T}$), two three-grid complex shifted preconditioners ($\mathcal{M}_{3,V}$, $\mathcal{M}_{3,F}$) and a two-grid cycle with a two-grid complex shifted coarse preconditioner ($\mathcal{T}_{2,V}$), respectively. We have considered hierarchies with at most three grids to yield a reasonable problem size per core. As in Section 5.2, we have used relaxation parameters issued from the rigorous Fourier analysis (relation (16)).

Table IV collects the number of preconditioner applications (Prec), computational times (T) and maximal requested memory (M) for these variants. With respect to the two-grid cycle $\mathcal{T}$, the combined cycle $\mathcal{T}_{2,V}$ is found to require a reduced number of preconditioner applications. Indeed, if we consider the case of $f = 20$ Hz, we remark a significant reduction of preconditioner applications when comparing the two-grid preconditioner $\mathcal{T}$ with the combined two-grid cycle $\mathcal{T}_{2,V}$ (248 versus 73). This also leads to a dramatic reduction of computational times (3346 s versus 748 s at $f = 20$ Hz). The $\mathcal{T}_{2,V}$ strategy always delivers the minimal computational times (see bold values in Table IV) among the four preconditioners with a clear advantage at medium to large frequencies. Nevertheless, we would like to stress that the shifted preconditioner presented in Algorithm 2 is based on a combination of standard multigrid components. It is most likely that the use of Galerkin coarse grid approximation or of operator-dependent transfer operators could be

Table IV. Preconditioned flexible methods for the solution of the Helmholtz equation for the heterogeneous velocity field EAGE/SEG salt dome. Case of a second-order discretization with 10 points per wavelength such that relation (5) is satisfied. Case of two-grid ($\mathcal{T}$), of complex shifted multigrid cycles ($\mathcal{M}_{3,V}$, $\mathcal{M}_{3,F}$) and of combined cycles ($\mathcal{T}_{2,V}$) applied as a preconditioner of FGMRES(5).

EAGE/SEG salt dome									
f	h	Grid	# Cores	Prec	T (s)	M (GB)	Prec	T (s)	M (GB)
				$\mathcal{T}$			$\mathcal{T}_{2,V}$		
2.5	60	$231 \times 231 \times 71$	4	12	146	0.6	11	**98**	0.6
5	30	$463 \times 463 \times 143$	32	25	316	4.5	16	**147**	4.6
10	15	$927 \times 927 \times 287$	256	71	927	35.9	28	**270**	36.6
20	7.5	$1855 \times 1855 \times 575$	2048	248	3346	288.1	73	**748**	293.8
40	3.75	$3711 \times 3711 \times 1149$	16384	1000†	13912	2304.1	283	**3101**	2349.9
				$\mathcal{M}_{3,V}$			$\mathcal{M}_{3,F}$		
2.5	60	$231 \times 231 \times 71$	4	98	132	0.5	122	193	0.5
5	30	$463 \times 463 \times 143$	32	217	300	3.8	184	298	3.8
10	15	$927 \times 927 \times 287$	256	445	638	30.5	334	561	30.5
20	7.5	$1855 \times 1855 \times 575$	2048	2485	4102	244.8	2149	3764	244.8
40	3.75	$3711 \times 3711 \times 1149$	16384	8000†	—	1957.8	8000†	—	1957.8

Prec, the number of preconditioner applications; T, the total computational time in seconds; M, the requested memory in GB.
The bold values represent the minimal computational times, which $\mathcal{T}_{2,V}$ strategy always delivers.
Numerical experiments were performed on an IBM BG/P computer.
†A maximal number of preconditioner applications has been reached.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

beneficial to improve the properties of the preconditioner when considering heterogeneous Helmholtz problems. Despite the simplicity of the shifted preconditioner, we remark that both $\mathcal{M}_{3,V}$ and $\mathcal{M}_{3,F}$ strategies are more attractive than the two-grid preconditioner $\mathcal{T}$ in terms of computational times at small to medium range frequencies (2.5, 5 and 10 Hz, respectively). However, at high frequencies (20 and 40 Hz) a significant increase in terms of preconditioning applications is observed for both $\mathcal{M}_{3,V}$ and $\mathcal{M}_{3,F}$. We also notice that a shifted preconditioner based on an F-cycle is preferable when large frequencies are considered, that is solving approximately the coarse problem twice in a given cycle is found to be beneficial to the outer convergence.

In Figure 7, we consider the case of $f = 10$ Hz and represent the Ritz and harmonic Ritz values collected at each cycle of FGMRES(5) during convergence. As shown in [20], this computation allows us to investigate the quality of the *variable* preconditioner at a cheap cost, and we refer the reader to [55] for the definition of Ritz and harmonic Ritz values in this setting. Interestingly, the $\mathcal{T}$, $\mathcal{M}_{3,V}$ and $\mathcal{M}_{3,F}$ preconditioners lead to several outliers or clusters located in specific parts of the complex plane (even in the vicinity of the origin), whereas all Ritz or harmonic Ritz values are located in the unit disc (reasonably away from the origin) for the $\mathcal{T}_{2,V}$ preconditioner. Finally, we note that the combined cycle $\mathcal{T}_{2,V}$ used as a preconditioner of FGMRES(5) is also efficient when solving the largest frequency case ($f = 40$ Hz). A moderate number of preconditioner applications (283) and a low memory requirement (about 2.3 TB) are required to solve approximately this truly challenging case. This can be considered as a very satisfactory result and proves the usefulness of the algorithm on this realistic test case.

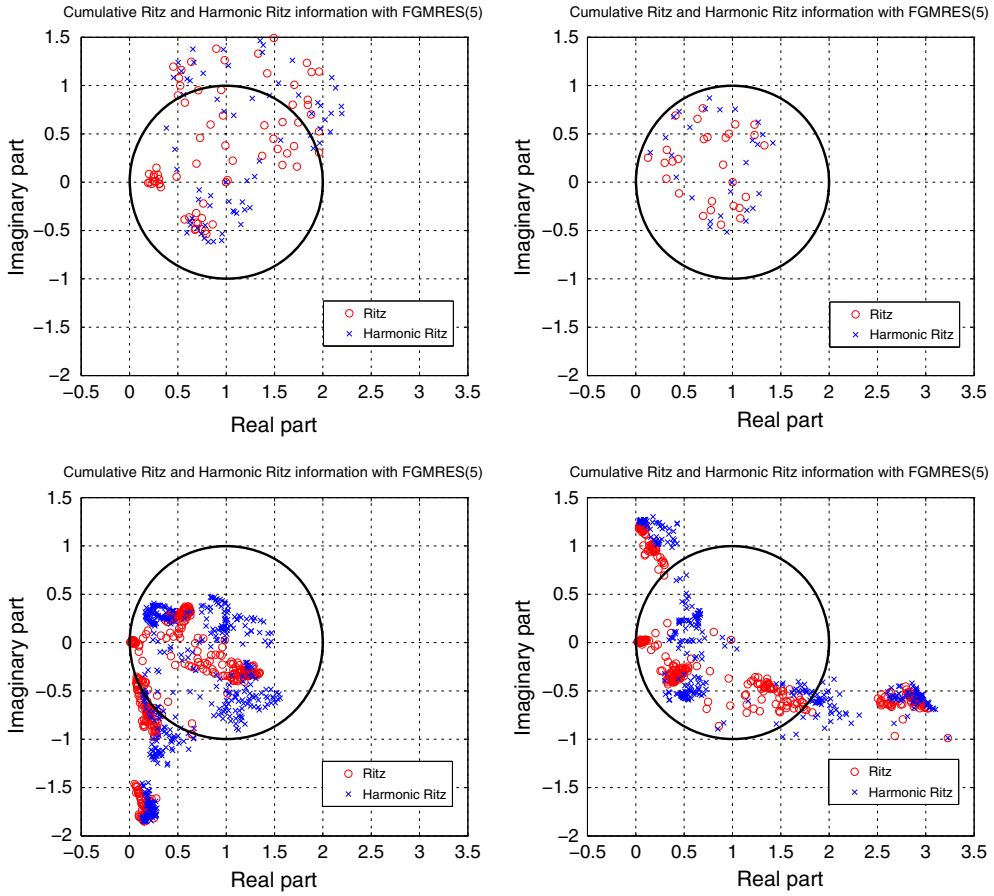


Figure 7. EAGE/SEG salt dome problem (case of $f = 10$ Hz, $927 \times 927 \times 287$ grid). Ritz and harmonic Ritz values (circles and crosses, respectively) of FGMRES(5) with four different variable preconditioners: $\mathcal{T}$ (top left), $\mathcal{T}_{2,V}$ (top right), $\mathcal{M}_{3,V}$ (bottom left) and $\mathcal{M}_{3,F}$ (bottom right) along convergence. Note that the same scales have been used for the four plots.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

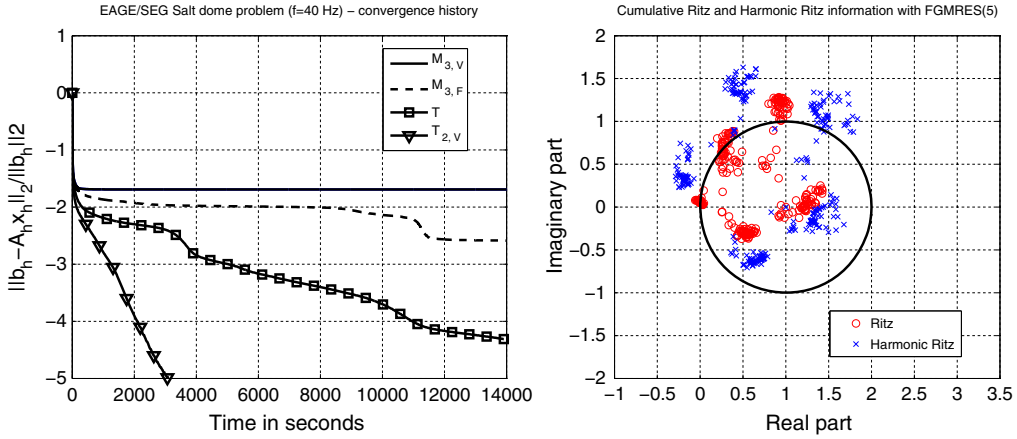


Figure 8. EAGE/SEG salt dome problem (case of $f = 40$ Hz, $3711 \times 3711 \times 1149$ grid). Convergence history of FGMRES(5) with four different variable preconditioners: $\mathcal{M}_{3,V}$ (line), $\mathcal{M}_{3,F}$ (dashed line), $\mathcal{T}$ (square) and $\mathcal{T}_{2,V}$ (triangle) versus computational times in seconds (left part). Ritz and harmonic Ritz values (circles and crosses, respectively) of FGMRES(5) with the $\mathcal{T}_{2,V}$ preconditioner (right part).

Remarks. We have also performed some numerical experiments with a larger value of the restarting parameter m in the outer Krylov subspace method FGMRES(m) ($m = 10$, results not shown here). At $f = 20$ Hz, a reduction of preconditioner applications is obtained for each strategy leading to a decrease of 10% in computational times. This slight improvement comes, however, at a price of increased memory requirements. Keeping memory consumption as low as possible is an important issue in this application, because we target the solution of multiple right-hand side problems with preconditioned block flexible Krylov subspace methods as discussed in [22]. Hence, we have preferred to focus on preconditioned FGMRES(5) in this section and to show the related performance even for such moderate value of the restarting parameter m. We refer the reader to [42] for a theoretical analysis of inner–outer methods when the outer and the inner methods are the same (FGMRES and GMRES in our setting). It is notably proved that by using preconditioners, which are Krylov methods, the global iteration is maintained within a larger Krylov subspace.

Figure 8 (left part) shows the convergence history of FGMRES(5) with four different preconditioners, namely $\mathcal{T}$, $\mathcal{M}_{3,V}$, $\mathcal{M}_{3,F}$ and $\mathcal{T}_{2,V}$ on the most challenging case ($f = 40$ Hz, approximately 15.8 billions of unknowns). Interestingly, we notice that the stopping criterion (27) is satisfied only for FGMRES(5) used in combination with the new preconditioner $\mathcal{T}_{2,V}$ (see right part of Figure 8 for the repartition of Ritz and harmonic Ritz values). The $\mathcal{T}$, $\mathcal{M}_{3,V}$ and $\mathcal{M}_{3,F}$ approaches lead to a certain residual reduction, but due to limited computing resources, we have only reported the maximal number of preconditioner applications and related elapsed computational times in Table IV. We remark that a long-term stagnation in the convergence does appear for the shifted preconditioner. We further plan to analyse this behaviour in the light of recent non-stagnation conditions for the convergence of GMRES on indefinite problems [56, 57] as part of future work. Contrary to the case of $f = 20$ Hz (Figure 7, top right part) Ritz and harmonic Ritz with negative real part are observed at $f = 40$ Hz (Figure 8, right part) for the $\mathcal{T}_{2,V}$ combined preconditioner.[‖] We refer the reader to comments given in the last paragraph of this section for possible improvements related to both preconditioner and outer Krylov subspace method.

We report numerical results in Table V related to the $\mathcal{T}$ and $\mathcal{T}_{2,V}$ cycles, now with a different smoothing strategy at the fine level only. A fine level smoother based on either one cycle of unpreconditioned GMRES(4) ($\vartheta = 1$, $m_s = 4$ and $\nu = 0$) or two cycles of unpreconditioned GMRES(2) ($\vartheta = 2$, $m_s = 2$ and $\nu = 0$) is investigated (without any changes for the other parameters). This choice leads to the same number of matrix–vector products to be performed in the smoother on the

[‖]More precisely, 25 Ritz values with negative real part (smallest modulus equal to 0.05) and 56 harmonic Ritz values with negative real part (smallest modulus equal to 0.29) are obtained.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

Table V. Preconditioned flexible methods for the solution of the Helmholtz equation for the heterogeneous velocity field EAGE/SEG salt dome. Case of a second-order discretization with 10 points per wavelength such that relation (5) is satisfied. Case of two-grid cycles applied as a preconditioner of FGMRES(5).

EAGE/SEG salt dome								
f	Grid	# Cores	Prec	T (s)	M (GB)	Prec	T (s)	M (GB)
			$\mathcal{T}$ $(\vartheta = 2, m_s = 2, \nu = 0)$			$\mathcal{T}$ $(\vartheta = 1, m_s = 4, \nu = 0)$		
2.5	$231 \times 231 \times 71$	4	11	77	0.6	11	76	0.6
5	$463 \times 463 \times 143$	32	24	174	4.5	24	172	4.9
10	$927 \times 927 \times 287$	256	54	409	35.9	54	402	39.8
20	$1855 \times 1855 \times 575$	2048	180	1420	288.1	178	1381	319.5
40	$3711 \times 3711 \times 1149$	16384	1343	11180	2304.1	1244	10169	2444.7
			$\mathcal{T}_{2,V}$ $(\vartheta = 2, m_s = 2, \nu = 0)$			$\mathcal{T}_{2,V}$ $(\vartheta = 1, m_s = 4, \nu = 0)$		
2.5	$230 \times 231 \times 71$	4	11	89	0.6	11	88	0.9
5	$463 \times 463 \times 143$	32	16	135	4.6	16	133	6.0
10	$927 \times 927 \times 287$	256	28	247	36.6	27	235	43.0
20	$1855 \times 1855 \times 575$	2048	72	683	293.8	72	672	325.2
40	$3711 \times 3711 \times 1149$	16384	313	3929	2349.9	313	3885	2522.1

Prec, the number of preconditioner applications; T, the total computational time in seconds; M, the requested memory in GB.
Numerical experiments were performed on a IBM BG/P computer. Influence of the fine level smoother.

fine level as the initial setting ($\vartheta = 1$, $m_s = 2$ and $\nu = 2$). Interestingly, we note a significant reduction in terms of preconditioner applications for medium to large frequencies for the two-grid preconditioner $\mathcal{T}$. For this approach, performing more smoothing iterations by either restarting or increasing the degree of the polynomial smoother is then found to be beneficial on this given application. For the combined cycle $\mathcal{T}_{2,V}$, the change of the fine level smoother does not modify the number of preconditioner applications and leads to a reduction of at least 10% in terms of computational times for frequencies up to 20 Hz (see Table IV for a comparison). At $f = 40$ Hz, the cycle with preconditioned GMRES(2) ($\vartheta = 1$, $m_s = 2$ and $\nu = 2$) leads to better results in terms of preconditioner applications and computational times. Finally, we note that optimizing the sparse matrix–vector products [58] and considering communication avoiding GMRES method [59] in both the inner and outer Krylov subspace methods are two features that would be worth investigating to further reduce the computational times.

We have on purpose restricted our setting to simple multigrid components to be able to perform a rigorous Fourier analysis. Nevertheless, we are aware of possible improvements in the proposed algorithms. Indeed, smoothers based on symmetric Gauss–Seidel preconditioned GMRES (as studied in [20]), the use of Galerkin coarse grid approximation or of complex-valued operator-dependent transfer operators [53] might probably be beneficial to the preconditioners on heterogeneous problems. Moreover, given a certain preconditioner, considering the role of the flexible Krylov subspace method is certainly an issue to address in a near future. Other flexible methods [60, 61] or recent algorithms that include spectral information to improve the convergence rate—FGMRES-DR [55] or FGCRO-DR [62]—are definitively of interest in both inner and outer parts of the solver.

6. CONCLUSIONS

We have proposed a new two-grid preconditioner for the solution of Helmholtz problems in three-dimensional heterogeneous media. This two-grid cycle is applied directly to the Helmholtz operator and relies on an approximate coarse grid solution. A second multigrid method applied to a complex shifted Laplacian operator is then used as a preconditioner for the approximate solution of this coarse problem. Next, we have studied the convergence properties of this preconditioner with rigorous Fourier analysis and selected appropriate relaxation parameters for the smoother based on this

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

analysis. Finally we have highlighted the efficiency of the new preconditioner on both academic and concrete applications in geophysics requiring the solution of indefinite problems of huge dimension. Numerical results have demonstrated the usefulness of the combined algorithm on a realistic three-dimensional application at high frequency.

As part of future research, we plan to perform a three-grid Fourier analysis of the combined cycle to yield additional valuable insight into the preconditioning properties of this method. We will also investigate the behaviour of the combined preconditioner on problems issued from the high-order finite difference discretization of the acoustic or elastic Helmholtz equation [63] in both single and multiple source situations. To conclude, we note that the framework of the combined cycle can be extended to a fully algebraic setting by using algebraic multigrid ideas [37, Appendix A] (see also [64] for a specific extension to complex-valued problems) to construct the different operators involved in the two hierarchies. This may be especially useful when finite element discretizations of the Helmholtz equation (based, e.g. on Discontinuous Galerkin methods or on hp-finite element techniques) are considered. This is part of future research.

ACKNOWLEDGEMENTS

The authors would like to thank the two referees for their valuable comments and suggestions that helped us to improve the manuscript. They thank Prof. A. Borzi and Prof. C. W. Oosterlee for the invitation to the OPTPDE ESFWaves workshop held in Würzburg, Germany on September 26-28th 2011. They also would like to acknowledge GENCI (Grand Equipement National de Calcul Intensif) for the dotation of computing hours on the IBM BG/P computer at IDRIS, France. This work was granted access to the HPC resources of CINES and IDRIS under allocation 2011065068 and 2012065068 made by GENCI.

REFERENCES

1. Brandt A, Livshits I. Wave-ray multigrid method for standing wave equations. *Electronic Transactions on Numerical Analysis* 1997; **6**:162–181.
2. Brandt A, Ta'asan S. Multigrid method for nearly singular and slightly indefinite problems. In *Multigrid Methods II*, Hackbusch W, Trottenberg U (eds). Springer: Berlin Heidelberg, 1986; 99–121.
3. Farhat C, Macedo A, Lesoinne M. A two-level domain decomposition method for the iterative solution of high frequency exterior Helmholtz problems. *Numerische Mathematik* 2000; **85**:283–308.
4. Farhat C, Roux FX. A method of finite element tearing and interconnecting and its parallel solution algorithm. *International Journal for Numerical Methods in Engineering* 1991; **32**:1205–1227.
5. Toselli A, Widlund O. *Domain Decomposition Methods - Algorithms and Theory*, Springer Series on Computational Mathematics, Vol. 34. Springer: Berlin Heidelberg, 2005.
6. Erlangga YA. Advances in iterative methods and preconditioners for the Helmholtz equation. *Archives of Computational Methods in Engineering* 2008; **15**:37–66.
7. Ernst O, Gander MJ. Why it is difficult to solve Helmholtz problems with classical iterative methods. In *Numerical Analysis of Multiscale Problems*, Graham I, Lakkis O, Hou T, Scheichl R (eds). Springer: Berlin Heidelberg, 2011; 325–361.
8. Bayliss A, Goldstein CI, Turkel E. An iterative method for the Helmholtz equation. *Journal of Computational Physics* 1983; **49**:443–457.
9. Laird AL, Giles MB. Preconditioned iterative solution of the 2D Helmholtz equation. *Technical Report Report NA-02/12*, Oxford University Computing Laboratory, University of Oxford, United Kingdom, 2002.
10. Erlangga YA. A robust and efficient iterative method for the numerical solution of the Helmholtz equation. *Ph.D. Thesis*, TU Delft, Delft, The Netherlands, 2005.
11. Erlangga YA, Vuik C, Oosterlee C. On a class of preconditioners for solving the Helmholtz equation. *Applied Numerical Mathematics* 2004; **50**:409–425.
12. Erlangga YA, Oosterlee C, Vuik C. A novel multigrid based preconditioner for heterogeneous Helmholtz problems. *SIAM Journal on Scientific Computing* 2006; **27**:1471–1492.
13. van Gijzen MB, Erlangga YA, Vuik C. Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian. *SIAM Journal on Scientific Computing* 2007; **29**:1942–1958.
14. Riyanti CD, Kononov A, Erlangga YA, Plessix R-E, Mulder WA, Vuik C, Oosterlee C. A parallel multigrid-based preconditioner for the 3D heterogeneous high-frequency Helmholtz equation. *Journal of Computational Physics* 2007; **224**:431–448.
15. Bollhöfer M, Grote MJ, Schenk O. Algebraic multilevel preconditioner for the solution of the Helmholtz equation in heterogeneous media. *SIAM Journal on Scientific Computing* 2009; **31**:3781–3805.
16. Erlangga YA, Nabben R. On a multilevel Krylov method for the Helmholtz equation preconditioned by shifted Laplacian. *Electronic Transactions on Numerical Analysis* 2008; **31**:403–424.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

17. Sheikh AH, Lahaye D, Vuik C. A scalable Hemholtz solver combining the shifted Laplace preconditioner with multigrid deflation. *Report 11-01*, Delft University of Technology, Delft Institute of Applied Mathematics, Delft, 2011.
18. Elman H, Ernst O, O'Leary D, Stewart M. Efficient iterative algorithms for the stochastic finite element method with application to acoustic scattering. *Computer Methods in Applied Mechanics and Engineering* 2005; **194**(1):1037–1055.
19. Elman HC, Ernst OG, O'Leary DP. A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations. *SIAM Journal on Scientific Computing* 2001; **23**:1291–1315.
20. Pinel X. A perturbed two-level preconditioner for the solution of three-dimensional heterogeneous Helmholtz problems with applications to geophysics. *Ph.D. Thesis*, CERFACS, Toulouse, France, 2010. TH/PA/10/55, available at http://www.cerfacs.fr/algor/reports/Dissertations/TH_PA_10_55.pdf [Accessed date on 2010].
21. Stüben K, Trottenberg U. Multigrid methods: fundamental algorithms, model problem analysis and applications. In *Multigrid Methods, Koeln-Porz, 1981, Lecture Notes in Mathematics, Volume 960*, Hackbusch W, Trottenberg U (eds). Springer: Berlin Heidelberg, 1982; 1–176.
22. Calandra H, Gratton S, Langou J, Pinel X, Vasseur X. Flexible variants of block restarted GMRES methods with application to geophysics. *SIAM Journal on Scientific Computing* 2012; **34**(2):A714–A736.
23. Calandra H, Gratton S, Lago R, Pinel X, Vasseur X. Two-level preconditioned Krylov subspace methods for the solution of three-dimensional heterogeneous Helmholtz problems in seismics. *Numerical Analysis and Applications* 2012; **5**:175–181.
24. Virieux J, Operto S. An overview of full waveform inversion in exploration geophysics. *Geophysics* 2009; **74**(6):WCC127–WCC152.
25. Tarantola A. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM: Philadelphia, 2005.
26. Berenger J-P. A perfectly matched layer for absorption of electromagnetic waves. *Journal of Computational Physics* 1994; **114**:185–200.
27. Berenger J-P. Three-dimensional perfectly matched layer for absorption of electromagnetic waves. *Journal of Computational Physics* 1996; **127**:363–379.
28. Operto S, Virieux J, Amestoy PR, L'Excellent J-Y, Giraud L, Ben-Hadj-Ali H. 3D finite-difference frequency-domain modeling of visco-acoustic wave propagation using a massively parallel direct solver: a feasibility study. *Geophysics* 2007; **72-5**:195–211.
29. Cohen G. *Higher-order Numerical Methods for Transient Wave Equations*. Springer: Berlin Heidelberg, 2002.
30. Sourbier F, Operto S, Virieux J, Amestoy P, L'Excellent JY. FWT2D : a massively parallel program for frequency-domain full-waveform tomography of wide-aperture seismic data - part 1: algorithm. *Computer & Geosciences* 2009; **35**:487–495.
31. Sourbier F, Operto S, Virieux J, Amestoy P, L'Excellent JY. FWT2D : a massively parallel program for frequency-domain full-waveform tomography of wide-aperture seismic data - part 2: numerical examples and scalability analysis. *Computer & Geosciences* 2009; **35**:496–514.
32. Wang S, de Hoop MV, Xia J. Acoustic inverse scattering via Helmholtz operator factorization and optimization. *Journal of Computational Physics* 2010; **229**:8445–8462.
33. Wang S, de Hoop MV, Xia J. On 3D modeling of seismic wave propagation via a structured parallel multifrontal direct Helmholtz solver. *Geophysical Prospecting* 2011; **59**:857–873.
34. Engquist B, Ying L. Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers. *Multiscale Modeling and Simulation* 2011; **9**:686–710.
35. Reps B, Vanroose W, bin Zubair H. On the indefinite Helmholtz equation: complex stretched absorbing boundary layers, iterative analysis, and preconditioning. *Journal of Computational Physics* 2010; **229**:8384–8405.
36. Virieux J, Operto S, Ben-Hadj-Ali H, Brossier R, Etienne V, Sourbier F, Giraud L, Haidar A. Seismic wave modeling for seismic imaging. *The Leading Edge* 2009; **25**(8):538–544.
37. Trottenberg U, Oosterlee CW, Schüller A. *Multigrid*. Academic Press Inc.: San Diego, 2001.
38. Vanroose W, Reps B, bin Zubair H. A polynomial multigrid smoother for the iterative solution of the heterogeneous Helmholtz problem. *Technical Report*, University of Antwerp, Belgium, 2010. http://arxiv.org/abs/1012.5379 [Accessed date on 2010].
39. Saad Y, Schultz MH. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing* 1986; **7**:856–869.
40. De Zeeuw PM. Matrix-dependent prolongations and restrictions in a blackbox multigrid solver. *Journal of Computational and Applied Mathematics* 1990; **33**:1–27.
41. Notay Y, Vassilevski PS. Recursive Krylov-based multigrid cycles. *Numerical Linear Algebra with Applications* 2008; **15**:473–487.
42. Simoncini V, Szyld DB. Flexible inner-outer Krylov subspace methods. *SIAM Journal on Numerical Analysis* 2003; **40**:2219–2239.
43. Simoncini V, Szyld DB. Recent computational developments in Krylov subspace methods for linear systems. *Numerical Linear Algebra with Applications* 2007; **14**:1–59.
44. Vassilevski PS. *Multilevel Block Factorization Preconditioners, Matrix-Based Analysis and Algorithms for Solving Finite Element Equations*. Springer: New York, 2008.
45. Saad Y. A flexible inner–outer preconditioned GMRES algorithm. *SIAM Journal on Scientific and Statistical Computing* 1993; **14**:461–469.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

46. Thole CA, Trottenberg U. Basic smoothing procedures for the multigrid treatment of elliptic 3D operators. *Applied Mathematics and Computation* 1986; **19**:333–345.
47. Wienands R, Oosterlee CW, Washio T. Fourier analysis of GMRES(m) preconditioned by multigrid. *SIAM Journal on Scientific Computing* 2000; **22**:582–603.
48. Cools S, Vanroose W. Local Fourier analysis of the complex shifted Laplacian preconditioner for Helmholtz problems. *Technical report*, University of Antwerp, Department of Mathematics and Computer Science, Belgium, 2011.
49. Duff IS, Gratton S, Pinel X, Vasseur X. Multigrid based preconditioners for the numerical solution of two-dimensional heterogeneous problems in geophysics. *International Journal of Computer Mathematics* 2007; **84-88**:1167–1181.
50. Saad Y. *Iterative Methods for Sparse Linear Systems*, 2nd ed. SIAM: Philadelphia, 2003.
51. Hackbusch W, Trottenberg U. *Multigrid Methods*. Springer: Berlin Heidelberg, 1982. Lecture notes in Mathematics, Vol. 960, Proceedings of the conference held at Köln-Porz, November 23-27 1981.
52. Gropp W, Lusk E, Skjellum A. *Using MPI: Portable Parallel Programming with the Message-Passing Interface*. MIT Press: Cambridge, Massachusetts, 1999.
53. Umetani N, MacLachlan SP, Oosterlee CW. A multigrid-based shifted Laplacian preconditioner for fourth-order Helmholtz discretization. *Numerical Linear Algebra with Applications* 2009; **16**:603–626.
54. Aminzadeh F, Brac J, Kunz T. 3D salt and overthrust models. *SEG/EAGE modeling series I*, Society of Exploration Geophysicists, Tulsa, USA, 1997.
55. Giraud L, Gratton S, Pinel X, Vasseur X. Flexible GMRES with deflated restarting. *SIAM Journal on Scientific Computing* 2010; **32**:1858–1878.
56. Simoncini V. On a non-stagnation condition for GMRES and application to saddle point matrices. *Electronic Transactions on Numerical Analysis* 2010; **37**:202–213.
57. Simoncini V, Szyld DB. New conditions for non-stagnation of minimal residual methods. *Numerische Mathematik* 2008; **109**:477–487.
58. Datta K, Kamil S, Williams S, Oliker L, Shalf J, Yelick K. Optimization and performance modeling of stencil computations on modern microprocessors. *SIAM Review* 2009; **51**(1):129–159.
59. Hoemmen M. Communication-avoiding Krylov subspace methods. *Ph.D. Thesis*, University of California, Berkeley, Department of Computer Science, 2010.
60. Szyld DB, Vogel JA. FQMR: a flexible quasi-minimal residual method with inexact preconditioning. *SIAM Journal on Scientific Computing* 2001; **23**(2):363–380.
61. Vogel JA. Flexible BiCG and flexible Bi-CGSTAB for nonsymmetric linear systems. *Applied Mathematics and Computation* 2007; **188**:226–233.
62. Carvalho LM, Gratton S, Lago R, Vasseur X. A flexible generalized conjugate residual method with inner orthogonalization and deflated restarting. *SIAM Journal on Matrix Analysis and Applications* 2011; **32**(4):1212–1235.
63. Harari I, Turkel E. Accurate finite difference methods for time-harmonic wave propagation. *Journal of Computational Physics* 1995; **119**:252–270.
64. MacLachlan SP, Oosterlee CW. Algebraic multigrid solvers for complex-valued matrices. *SIAM Journal on Scientific Computing* 2008; **30**:1548–1571.

Copyright © 2012 John Wiley & Sons, Ltd.

DOI: 10.1002/nla

B.5. A modified block flexible GMRES method with deflation at each iteration for the solution of non-Hermitian linear systems with multiple right-hand sides

SIAM J. SCI. COMPUT.
Vol. 35, No. 5, pp. S345–S367
© 2013 Society for Industrial and Applied Mathematics

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

A MODIFIED BLOCK FLEXIBLE GMRES METHOD WITH DEFLATION AT EACH ITERATION FOR THE SOLUTION OF NON-HERMITIAN LINEAR SYSTEMS WITH MULTIPLE RIGHT-HAND SIDES*

HENRI CALANDRA†, SERGE GRATTON‡, RAFAEL LAGO§, XAVIER VASSEUR¶, AND LUIZ MARIANO CARVALHO‖

Abstract. We propose a variant of the block GMRES method for the solution of linear systems of equations with multiple right-hand sides. We investigate a deflation strategy to detect when a linear combination of approximate solutions is already known that avoids performing expensive computational operations with the system matrix. This is especially useful when the cost of the preconditioner is supposed to be larger than the cost of orthogonalization in the block Arnoldi procedure. We specifically focus on the block GMRES method incorporating deflation at the end of each iteration proposed by Robbé and Sadkane [M. Robbé and M. Sadkane, *Linear Algebra Appl.*, 419 (2006), pp. 265–285]. We extend their contribution by proposing that deflation be performed also at the beginning of each cycle. This change leads to a modified least-squares problem to be solved at each iteration and gives rise to a different behavior especially when multiple restarts are required to reach convergence. Additionally we investigate truncation techniques, aiming at reducing the computational cost of the iteration. This is particularly useful when the number of right-hand sides is large. Finally, we address the case of variable preconditioning, an important feature when iterative methods are used as preconditioners, as investigated here. The numerical experiments performed in a parallel environment show the relevance of the proposed variant on a challenging application related to geophysics. A savings of up to 35% in terms of computational time—at the same memory cost—is obtained with respect to the original method on this application.

Key words. block Krylov space method, block size reduction, deflation at each iteration, flexible preconditioning, multiple right-hand sides

AMS subject classifications. 65F10, 65N22, 15A06

DOI. 10.1137/120883037

1. Introduction.

We consider block Krylov space methods for the solution of linear systems of equations with p right-hand sides given at once of the form $AX = B$, where $A \in \mathbb{C}^{n\times n}$ is supposed to be a nonsingular non-Hermitian matrix, $B \in \mathbb{C}^{n\times p}$ is supposed to be full rank, and $X \in \mathbb{C}^{n\times p}$. Although the number of right-hand sides p might be relatively large, we suppose here that the dimension of the problem n is always much larger. Later, we denote by $X_0 \in \mathbb{C}^{n\times p}$ the initial block iterate, and by $R_0 = B - AX_0$ the initial block residual. As stated in [25, 26] a block Krylov space method for solving the p systems is an iterative method that generates approximations $X_m \in \mathbb{C}^{n\times p}$ with $m \in \mathbb{N}$ such that

$$X_m - X_0 \in \mathcal{K}_m^{\square}(A, R_0),$$

*Received by the editors June 29, 2012; accepted for publication (in revised form) June 25, 2013; published electronically October 28, 2013.
http://www.siam.org/journals/sisc/35-5/88303.html
†TOTAL, Centre Scientifique et Technique Jean Féger, F-64000 Pau, France (Henri.Calandra@total.com).
‡INPT-IRIT, University of Toulouse, and ENSEEIHT, BP 7122, F-31071 Toulouse Cedex 7, France (Serge.Gratton@enseeiht.fr).
§CERFACS, F-31057 Toulouse Cedex 1, France (lago@cerfacs.fr).
¶CERFACS and HiePACS project joint INRIA-CERFACS Laboratory, F-31057 Toulouse Cedex 1, France (vasseur@cerfacs.fr).
‖Applied Mathematics Department, IME-UERJ, 20559-900, Rio de Janeiro, RJ, Brazil (luizmc@gmail.com).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

where the block Krylov space $K_m^{\square}(A, R_0)$ (in the nonpreconditioned case) is defined as

$$\mathcal{K}_m^{\square}(A, R_0) = \left\{ \sum_{k=0}^{m-1} A^k R_0 \gamma_k \ \forall \gamma_k \in \mathbb{C}^{p\times p}, \text{ with } k \mid 0 \le k \le m-1 \right\} \subset \mathbb{C}^{n\times p}.$$

We refer the reader to [25] for a recent detailed overview on block Krylov subspace methods and note that most of the standard Krylov subspace methods have a block counterpart (see, e.g., block GMRES [48], block BiCGStab [24], block IDR(s) [16], and block QMR [22]). In this paper we mainly focus on restarted block Krylov subspace methods that satisfy a minimum norm property as introduced in [42, section 6.12].

Block Krylov subspace methods are increasingly popular in many application areas in computational science and engineering (e.g., electromagnetic scattering (monostatic radar cross section analysis) [10, 31, 44], lattice quantum chromodynamics [43], model reduction in circuit simulation [21], stochastic finite element with uncertainty restricted to the right-hand side [18], and sensitivity analysis of mechanical systems [7]). To be effective in terms of computational operations it is recognized that these methods must incorporate a strategy for detecting when a linear combination of the systems has approximately converged [25]. This explicit block size reduction is called deflation, as discussed in [25]. First, a simple strategy to remove useless information from a block Krylov subspace—called initial deflation—consists of detecting possible linear dependency in the block right-hand side B or in the initial block residual R_0 [25, section 12], [31, section 3.7.2]. When a restarted block Krylov subspace method is used, this block size reduction can be also performed at each initial computation of the block residual, i.e., at the beginning of each cycle [25, section 14]. In addition Arnoldi deflation [25] may be considered; it aims at detecting a near rank deficiency occurring in the block Arnoldi procedure to later reduce the current block size. These three strategies based on rank-revealing QR-factorizations [11] or singular value decompositions [23] have been notably proposed both in the Hermitian [35, 40] and non-Hermitian [1, 4, 14, 22, 33, 36] cases for block Lanczos methods. They have been shown to be effective with respect to standard block Krylov subspace methods. While initial deflation or deflation at the beginning of a cycle are currently popular, block Krylov subspace methods based on a norm minimization property incorporating deflation at each iteration have rarely been studied (see, e.g., [7] for a discussion on deflated block Arnoldi methods).

In this paper we focus only on block GMRES based methods [48] and refer the reader to [7, 22, 33, 34] for advanced block Lanczos methods with deflation. In [39] Robbé and Sadkane introduced the notion of inexact breakdown to study block size reduction techniques in block GMRES. Two criteria have been proposed, based either on the numerical rank of the generated block Krylov basis (W-criterion) or on the numerical rank of the block residual (R-criterion). Numerical experiments on academic problems of small dimension with a reduced number of right-hand sides illustrated the advantages and drawbacks of each variant versus standard block GMRES. Further numerical experiments can be found in [29]. Another method relying on such a strategy is the dynamic BGMRES (DBGMRES) [15], which is an extension of block loose GMRES [5]. We also refer the reader to [7], where deflated block Arnoldi methods, in addition to Lanczos, are discussed on a real application problem in structural mechanics. The combination of block GMRES performing deflation at each iteration and variable preconditioning has rarely been addressed in the literature. Variable preconditioning is often required when solving large linear systems of equations. This

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

is notably the case when inexact solutions of the preconditioning system using, e.g., nonlinear smoothers in multigrid [37] or approximate interior solvers in domain decomposition methods [46, section 4.3] are considered. Thus the main purpose of the paper is to derive a flexible minimal norm block Krylov subspace method that incorporates block size reduction at each iteration suited to the solution of large-scale linear systems (where expensive variable preconditioners are often used) with possibly a large number of right-hand sides. This is especially useful when the cost of the preconditioner is supposed to be larger than the cost of orthogonalization in the block Arnoldi procedure.

The paper is organized as follows. First we will introduce in section 2 the block GMRES method with deflation at each iteration proposed in [39], since it will constitute the basis for further developments. We will notably describe how deflation at each iteration is performed. In section 3 we first explain the main motivations for deriving the proposed variant and analyze its main mathematical properties. Algorithmic details are then presented in section 4 together with an analysis of the computational cost and memory requirements. Then in section 5 we demonstrate the effectiveness of the proposed algorithm on an application related to geophysics. Finally, we draw some conclusions in section 6.

2. Block GMRES with deflation at each iteration. In this section we review the block GMRES method with deflation at each iteration (later denoted BGMRES-R[1]) [39] for the solution of linear systems with a non-Hermitian matrix and multiple right-hand sides given at once. We first introduce notation used in the manuscript and then describe the main mathematical properties of BGMRES-R.

2.1. Notation. Throughout this paper we denote by $\|.\|_2$ the Euclidean norm, by $\|.\|_F$ the Frobenius norm, by $I_k \in \mathbb{C}^{k\times k}$ the identity matrix of dimension k, and by $0_{i\times j} \in \mathbb{C}^{i\times j}$ the zero rectangular matrix with i rows and j columns. The superscript ^{H} denotes the transpose conjugate operation. Given a vector $d \in \mathbb{C}^k$ with components d_i, $D = \mathrm{diag}(d_1, \ldots, d_k)$ is the diagonal matrix $D \in \mathbb{C}^{k\times k}$ such that $D_{ii} = d_i$. If $C \in \mathbb{C}^{k\times l}$, we denote the singular values of C by $\sigma_1(C) \geq \cdots \geq \sigma_{\min(k,l)}(C) \geq 0$. Finally, $e_m \in \mathbb{C}^n$ denotes the mth canonical basis vector of $\mathbb{C}^n$. In describing our algorithms (Algorithms 1–4), we adopt notation similar to that of MATLAB. For instance, $U(i, j)$ denotes the U_{ij} entry of matrix U, $U(1 : m, 1 : j)$ refers to the submatrix made of the first m rows and first j columns of U, and $U(:, j)$ corresponds to its jth column.

2.2. Overview. Next we provide a brief overview of the block GMRES method with deflation at each iteration, introduced in [39], and specifically focus on the variant with a block size reduction strategy based on the numerical rank of the block residual (R-criterion [39, section 4]). More precisely we propose to analyze a given cycle of this method in the next subsections. Throughout the paper we denote by $X_0 \in \mathbb{C}^{n\times p}$ the current approximation of the solution, and by $R_0 \in \mathbb{C}^{n\times p}$ the corresponding true block residual ($R_0 = B - AX_0$), both obtained at the beginning of the cycle that we consider. $D \in \mathbb{C}^{p\times p}$ represents a nonsingular diagonal scaling matrix defined as $D = \mathrm{diag}(b_1, \ldots, b_p)$ with $b_l = ||B(:, l)||_2$, $1 \leq l \leq p$. Finally, we assume that the QR factorization of R_0D^{-1} has been performed as

$$R_0D^{-1} = \hat{\mathcal{V}}_1\hat{\Lambda}_0, \tag{2.1}$$

[1]The suffix "R" is used to emphasize that we exclusively consider the block GMRES method with deflation at each iteration based on the R-criterion proposed by Robbé and Sadkane in [39].

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

with $\hat{\mathcal{V}}_1 \in \mathbb{C}^{n\times p}$ having orthonormal columns and $\hat{\Lambda}_0 \in \mathbb{C}^{p\times p}$ assuming[2] $\operatorname{rank}(R_0 D^{-1}) = p$. R_0 ($R_0 D^{-1}$) is named the initial block residual (respectively, scaled initial block residual), where the term "initial" refers to the beginning of the cycle that we consider.

2.2.1. Deflated Arnoldi relation. If $K \in \mathbb{C}^{n\times p}$ denotes a matrix with orthonormal columns containing all the p new Krylov directions at iteration $j-1$, the most expensive part of the algorithm at the jth iteration lies in the p applications of the variable preconditioner supposed to be expensive. To be effective in terms of computational operations it is widely recognized that block Krylov subspace methods must rely on a strategy for detecting when a linear combination of the systems has approximately converged [25, 31]. In the framework of block Krylov subspace methods based on a norm minimization property, Robbé and Sadkane [39] have first proposed a block GMRES algorithm that relies on deflation at each iteration of a given cycle. To do so, they have introduced a modified version of the block Arnoldi algorithm—later called *deflated block Arnoldi*—in which $\operatorname{range}(K)$ is judiciously decomposed into

$$\operatorname{range}(K) = \operatorname{range}(V_j) \oplus \operatorname{range}(P_{j-1}), \quad \text{with } \begin{bmatrix} V_j & P_{j-1} \end{bmatrix}^H \begin{bmatrix} V_j & P_{j-1} \end{bmatrix} = I_p, \tag{2.2}$$

where $V_j \in \mathbb{C}^{n\times k_j}$, $P_{j-1} \in \mathbb{C}^{n\times d_j}$ with $k_j + d_j = p$. In other words, k_j Krylov directions are effectively considered at iteration j, while d_j directions are left aside (or deflated) at the same iteration. We note that literally the "best" subspace of $\operatorname{range}(K)$ of dimension k_j is chosen (not just k_j columns of K) defining V_j, leaving the remaining subspace in $\operatorname{range}(P_{j-1})$ (i.e., the deflated subspace is spanned by $\operatorname{range}(P_{j-1})$ at iteration j). Based on this decomposition, the deflated orthonormalization procedure will apply preconditioning and matrix-vector products *only* over the chosen k_j directions of V_j. Next we briefly describe the jth iteration of the resulting method.

Defining $s_0 = 0$, $s_j = s_{j-1} + k_j$ and given $\begin{bmatrix} \mathcal{V}_j & P_{j-1} \end{bmatrix} \in \mathbb{C}^{n\times(s_j+d_j)}$ with orthonormal columns, the following block Arnoldi relation is assumed to hold at the beginning of the jth iteration of the deflated block Arnoldi procedure ($j > 1$):

$$A\mathcal{V}_{j-1} = \begin{bmatrix} \mathcal{V}_j & P_{j-1} \end{bmatrix} \mathcal{H}_{j-1}, \tag{2.3}$$

with $\mathcal{V}_{j-1} \in \mathbb{C}^{n\times s_{j-1}}$, $\mathcal{V}_j \in \mathbb{C}^{n\times s_j}$, $P_{j-1} \in \mathbb{C}^{n\times d_j}$, and $\mathcal{H}_{j-1} \in \mathbb{C}^{(s_{j-1}+p)\times s_{j-1}}$. The jth iteration of the deflated block Arnoldi procedure produces matrices $\hat{V}_{j+1} \in \mathbb{C}^{n\times k_j}$, $\hat{\mathcal{H}}_j \in \mathbb{C}^{(s_j+p)\times s_j}$ which satisfy

$$A \begin{bmatrix} \mathcal{V}_{j-1} & V_j \end{bmatrix} = \begin{bmatrix} \mathcal{V}_j & P_{j-1} & \hat{V}_{j+1} \end{bmatrix} \hat{\mathcal{H}}_j, \tag{2.4}$$

where $\hat{\mathcal{H}}_j$ has the following block structure:

$$\hat{\mathcal{H}}_j = \left[\begin{array}{c|c} \mathcal{H}_{j-1} & H_j \\ \hline 0_{k_j\times s_{j-1}} & H_{j+1,j} \end{array}\right],$$

with $H_j \in \mathbb{C}^{(s_{j-1}+p)\times k_j}$ and $H_{j+1,j} \in \mathbb{C}^{k_j\times k_j}$ (see Algorithm 1 for a complete description of this iteration). We assume that $\hat{\mathcal{H}}_j$ is always full rank; i.e., no Arnoldi breakdown occurs. We note that Arnoldi breakdowns rarely happen in practice (see, e.g.,

[2]The situation of $R_0 D^{-1}$ being rank-deficient in exact arithmetic is often referred to as *initial breakdown* [25]. However, as in [39], for the sake of simplicity we consider that $\operatorname{rank}(R_0 D^{-1}) = p$ holds at each cycle. We refer the reader to [25] for details on how to work around initial deflation, and we point out that this phenomenon has not been observed in our numerical experiments.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

[25, section 13]). Therefore the possibility of an Arnoldi breakdown has not been considered in this paper, as in recent contributions [13, 25, 39]. Defining $\hat{\mathcal{V}}_{j+1} \in \mathbb{C}^{n\times(s_j+p)}$ as

$$\hat{\mathcal{V}}_{j+1} = \begin{bmatrix} \mathcal{V}_j & P_{j-1} & \hat{V}_{j+1} \end{bmatrix}, \tag{2.5}$$

the block Arnoldi relation (2.4) can then be stated as

$$A\mathcal{V}_j = \hat{\mathcal{V}}_{j+1}\hat{\mathcal{H}}_j. \tag{2.6}$$

Next the key idea is to perform the subspace decomposition previously mentioned in (2.2) as

$$\begin{aligned} \begin{bmatrix} \mathcal{V}_j & V_{j+1} & P_j \end{bmatrix} &= \begin{bmatrix} \mathcal{V}_j & P_{j-1} & \hat{V}_{j+1} \end{bmatrix} \mathcal{F}_{j+1}, \\ \begin{bmatrix} \mathcal{V}_{j+1} & P_j \end{bmatrix} &= \hat{\mathcal{V}}_{j+1}\mathcal{F}_{j+1}, \end{aligned} \tag{2.7}$$

where $\mathcal{F}_{j+1} \in \mathbb{C}^{(s_j+p)\times(s_j+p)}$ is a unitary matrix. We address how to determine $\mathcal{F}_{j+1}$ later in section 2.2.4. Hence we obtain

$$A\mathcal{V}_j = \hat{\mathcal{V}}_{j+1}\mathcal{F}_{j+1}\mathcal{F}_{j+1}^H\hat{\mathcal{H}}_j.$$

Defining $\mathcal{H}_j \in \mathbb{C}^{(s_j+p)\times s_j}$ as $\mathcal{H}_j = \mathcal{F}_{j+1}^H\hat{\mathcal{H}}_j$, we then deduce (since $\mathcal{F}_{j+1}$ is unitary)

$$A\mathcal{V}_j = \begin{bmatrix} \mathcal{V}_{j+1} & P_j \end{bmatrix} \mathcal{H}_j,$$

which is precisely the block Arnoldi relation required at the beginning of the next iteration (compare with relation (2.3)). This last relation can be written as

$$A\mathcal{V}_j = \begin{bmatrix} \mathcal{V}_{j+1} & P_j \end{bmatrix} \begin{bmatrix} \mathcal{L}_j \\ G_j \end{bmatrix},$$

where $\mathcal{L}_j$ corresponds to the $(s_j + k_{j+1}) \times s_j$ upper part of $\mathcal{H}_j$, and G_j to the $d_{j+1} \times s_j$ lower part of $\mathcal{H}_j$. This is exactly the core relation proposed in [39, section 5, Algorithm 2].

2.2.2. Representation of the scaled initial block residual. At the beginning of the cycle the initial subspace decomposition is supposed to hold in BGMRES-R:

$$\mathcal{V}_1 = \hat{\mathcal{V}}_1. \tag{2.8}$$

Consequently p Krylov directions are effectively considered at the first iteration of a given cycle ($k_1 = p$), while no directions are deflated at the same iteration ($d_1 = 0$). At iteration j of the cycle ($1 \le j \le m$), we define the quantity $\hat{\Lambda}_j \in \mathbb{C}^{(s_j+p)\times p}$ as

$$\hat{\Lambda}_j = \begin{bmatrix} \hat{\Lambda}_0 \\ 0_{s_j\times p} \end{bmatrix}. \tag{2.9}$$

It is then straightforward to prove that R_0D^{-1} can be written as

$$R_0D^{-1} = \hat{\mathcal{V}}_{j+1}\hat{\Lambda}_j, \tag{2.10}$$

which means that $\hat{\Lambda}_j$ is the reduced representation of the scaled initial block residual in the $\hat{\mathcal{V}}_{j+1}$ basis.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

2.2.3. Minimization property. We denote by $Y_j \in \mathbb{C}^{s_j \times p}$ the solution of the reduced minimization problem $\mathcal{P}_r$ considered in BGMRES-R:

$$\mathcal{P}_r : Y_j = \underset{Y \in \mathbb{C}^{s_j \times p}}{\operatorname{argmin}} ||\hat{\Lambda}_j - \hat{\mathcal{H}}_j Y||_F, \tag{2.11}$$

with $\hat{\mathcal{H}}_j$ and $\hat{\Lambda}_j$ defined in (2.6) and (2.9), respectively. We also denote by $\hat{\mathcal{R}}_j \in \mathbb{C}^{(s_j+p)\times p}$ the block residual of the reduced least-squares problem $\mathcal{P}_r$, i.e., $\hat{\mathcal{R}}_j = \hat{\Lambda}_j - \hat{\mathcal{H}}_j Y_j$ $(1 \le j \le m)$, and define $\hat{\mathcal{R}}_0 \in \mathbb{C}^{p\times p}$ as $\hat{\mathcal{R}}_0 = \hat{\Lambda}_0$. We recall in Proposition 2.1 the norm minimization property occurring in BGMRES-R.

PROPOSITION 2.1. *In block GMRES with deflation at each iteration (BGMRES-R), solving the reduced minimization problem $\mathcal{P}_r$ of* (2.11) *amounts to minimizing the Frobenius norm of the block true residual $||B-AX||_F$ over the space $X_0+\text{range}(\mathcal{V}_j Y D)$ at iteration j $(1 \le j \le m)$ of a given cycle, i.e.,*

$$\begin{aligned} \underset{Y \in \mathbb{C}^{s_j \times p}}{\operatorname{argmin}} ||\hat{\Lambda}_j - \hat{\mathcal{H}}_j Y||_F &= \underset{Y \in \mathbb{C}^{s_j \times p}}{\operatorname{argmin}} ||R_0 D^{-1} - A\mathcal{V}_j Y||_F \\ &= \underset{Y \in \mathbb{C}^{s_j \times p}}{\operatorname{argmin}} ||B - A(X_0 + \mathcal{V}_j Y D)||_F, \end{aligned} \tag{2.12}$$

with $\hat{\mathcal{H}}_j$ and $\hat{\Lambda}_j$ defined in (2.6) *and* (2.9)*, respectively.*

Proof. Due to relations (2.4) and (2.10), $||R_0 D^{-1} - A\mathcal{V}_j Y||_F$ can be written as

$$||R_0 D^{-1} - A\mathcal{V}_j Y||_F = \|\hat{\mathcal{V}}_{j+1}(\hat{\Lambda}_j - \mathcal{H}_j Y)\|_F.$$

Since $\mathcal{V}_{j+1}$ has orthonormal columns and since the Frobenius norm is unitarily invariant, the last equality becomes

$$||R_0 D^{-1} - A\mathcal{V}_j Y||_F = \|\hat{\Lambda}_j - \hat{\mathcal{H}}_j Y\|_F.$$

D being a diagonal matrix, the relation (2.12) is then due to elementary properties of the Frobenius norm; namely, the squared Frobenius norm of a matrix is the sum of the squares of the Euclidean norms of its columns. ∎

2.2.4. Subspace decomposition based on a singular value decomposition. We next address the question of subspace decomposition; i.e., given $\hat{\mathcal{V}}_{j+1} = \begin{bmatrix}\mathcal{V}_j & \begin{bmatrix}P_{j-1} & \hat{V}_{j+1}\end{bmatrix}\end{bmatrix}$ obtained after the jth iteration of the deflated block Arnoldi procedure, we want to determine k_{j+1}, d_{j+1}, and the unitary matrix $\mathcal{F}_{j+1} \in \mathbb{C}^{(s_j+p)\times(s_j+p)}$ such that the decomposition (2.7) holds. To limit the computational cost related to the construction of $\mathcal{V}_{j+1}$, we consider the splitting $\mathcal{V}_{j+1} = \begin{bmatrix}\mathcal{V}_j & V_{j+1}\end{bmatrix}$ with $\mathcal{V}_j \in \mathbb{C}^{n\times s_j}$ obtained at the previous iteration and $V_{j+1} \in \mathbb{C}^{n\times k_{j+1}}$ to be determined. Thus the decomposition (2.7) can be written as

$$\begin{bmatrix}\mathcal{V}_j & \begin{bmatrix}V_{j+1} & P_j\end{bmatrix}\end{bmatrix} = \begin{bmatrix}\mathcal{V}_j & \begin{bmatrix}P_{j-1} & \hat{V}_{j+1}\end{bmatrix}\end{bmatrix} \mathcal{F}_{j+1}, \tag{2.13}$$

with $P_j \in \mathbb{C}^{n\times d_{j+1}}$ and $k_{j+1} + d_{j+1} = p$. Given the block form for $\mathcal{F}_{j+1}$,

$$\mathcal{F}_{j+1} = \begin{bmatrix}F_{11} & F_{12} \\ F_{21} & F_{22}\end{bmatrix},$$

where $F_{11} \in \mathbb{C}^{s_j \times s_j}$, $F_{12} \in \mathbb{C}^{s_j \times p}$, $F_{21} \in \mathbb{C}^{p\times s_j}$, and $F_{22} \in \mathbb{C}^{p\times p}$, the relation (2.13) becomes

$$\begin{bmatrix}\mathcal{V}_j & \begin{bmatrix}V_{j+1} & P_j\end{bmatrix}\end{bmatrix} = \begin{bmatrix}\mathcal{V}_j F_{11} + \begin{bmatrix}P_{j-1} & \hat{V}_{j+1}\end{bmatrix} F_{21} & \mathcal{V}_j F_{12} + \begin{bmatrix}P_{j-1} & \hat{V}_{j+1}\end{bmatrix} F_{22}\end{bmatrix}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Since $\mathcal{V}_j^H \begin{bmatrix} P_{j-1} & \hat{V}_{j+1} \end{bmatrix} = 0_{s_j \times p}$ we deduce the following matrix structure:

$$\mathcal{F}_{j+1} = \begin{bmatrix} I_{s_j} & 0_{s_j \times p} \\ 0_{p \times s_j} & F_j \end{bmatrix}, \tag{2.14}$$

where the unitary matrix $F_j \in \mathbb{C}^{p \times p}$ remains to be determined. The criterion proposed in [39] to deduce F_j, k_{j+1}, and d_{j+1} aims at finding a possible linear combination of the columns of $R_j D^{-1}$ that are approximately dependent (with respect to a certain threshold) to determine the set of directions that we do not want to consider when defining V_{j+1} in $\mathcal{V}_{j+1} = \begin{bmatrix} \mathcal{V}_j & V_{j+1} \end{bmatrix}$. Since $R_j D^{-1} = \hat{\mathcal{V}}_{j+1} \hat{\mathcal{R}}_j$, we instead perform this analysis based on the singular value decomposition of $\hat{\mathcal{R}}_j$ as $\hat{\mathcal{R}}_j = U \Sigma W^H$. We note that the thin singular value decomposition of $\hat{\mathcal{R}}_j$ is rather inexpensive since $\hat{\mathcal{R}}_j$ does not depend on the problem size n. Heuristically, *tol* being the convergence threshold used in the stopping criterion of BGMRES-R, we first choose a relative positive deflation threshold ε_d and then determine k_{j+1} according to the following condition:

$$\sigma_l(\hat{\mathcal{R}}_j) > \varepsilon_d \; tol \quad \forall \, l \, \text{such that} \, 1 \leq l \leq k_{j+1}. \tag{2.15}$$

Since $d_{j+1} = p - k_{j+1}$, the following decomposition of $\hat{\mathcal{R}}_j$ at iteration j is then obtained with $\hat{\mathcal{R}}_{s_j} \in \mathbb{C}^{s_j \times p}$ and $\hat{\mathcal{R}}_p \in \mathbb{C}^{p \times p}$:

$$\hat{\mathcal{R}}_j = \begin{bmatrix} \hat{\mathcal{R}}_{s_j} \\ \hat{\mathcal{R}}_p \end{bmatrix} = \begin{bmatrix} U_{s_j}^+ \\ U_p^+ \end{bmatrix} \Sigma_+ W_+^H + \begin{bmatrix} U_{s_j}^- \\ U_p^- \end{bmatrix} \Sigma_- W_-^H, \tag{2.16}$$

with $U_+ \in \mathbb{C}^{(s_j+p) \times k_{j+1}}$, $U_- \in \mathbb{C}^{(s_j+p) \times d_{j+1}}$, $\Sigma_+ \in \mathbb{C}^{k_{j+1} \times k_{j+1}}$, $\Sigma_- \in \mathbb{C}^{d_{j+1} \times d_{j+1}}$, $W_+ \in \mathbb{C}^{p \times k_{j+1}}$, and $W_- \in \mathbb{C}^{p \times d_{j+1}}$. Based on this splitting, Robbé and Sadkane have then proposed performing such a subspace decomposition at iteration j:

$$\text{range}((I_n - \mathcal{V}_j \mathcal{V}_j^H) R_j D^{-1}) = \text{range}(V_{j+1}) \oplus \text{range}(P_j),$$

where

$$\begin{aligned} \text{range}(V_{j+1}) &= \text{range}((I_n - \mathcal{V}_j \mathcal{V}_j^H) R_j D^{-1} W_+), \\ \text{range}(P_j) &= \text{range}((I_n - \mathcal{V}_j \mathcal{V}_j^H) R_j D^{-1} W_-), \end{aligned}$$

that is, the k_{j+1} directions associated with $(I_n - \mathcal{V}_j \mathcal{V}_j^H) R_j D^{-1} W_+$ (the kept ones) lie in V_{j+1}, while the d_{j+1} directions associated with $(I_n - \mathcal{V}_j \mathcal{V}_j^H) R_j D^{-1} W_-$ (the deflated ones, i.e., postponed and reintroduced later in next iterations if necessary) lie in P_j. Due to (2.16), this decomposition is also equivalent to

$$\begin{aligned} \text{range}(V_{j+1}) &= \text{range}\Big(\begin{bmatrix} P_{j-1} & \hat{V}_{j+1} \end{bmatrix} \begin{bmatrix} U_p^+ \end{bmatrix} \Sigma_+ \Big), \\ \text{range}(P_j) &= \text{range}\Big(\begin{bmatrix} P_{j-1} & \hat{V}_{j+1} \end{bmatrix} \begin{bmatrix} U_p^- \end{bmatrix} \Sigma_- \Big). \end{aligned}$$

Since $\begin{bmatrix} V_{j+1} & P_j \end{bmatrix} = \begin{bmatrix} P_{j-1} & \hat{V}_{j+1} \end{bmatrix} F_j$, the unitary matrix F_j is then simply obtained as the orthogonal factor of the QR decomposition of the $p \times p$ matrix $\begin{bmatrix} U_p^+ & U_p^- \end{bmatrix}$. This decomposition is summarized later in section 4, Algorithm 2.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

3. Modified block flexible GMRES with deflation at each iteration. In this section we present a modified block GMRES method with deflation at each iteration, which allows variable preconditioning and truncation, two features of significant interest when targeting the solution of large-scale non-Hermitian linear systems with possibly many right-hand sides. We first briefly introduce the motivations for these novelties and then describe the main mathematical properties of the resulting method, named BFGMRES-S.[3]

3.1. Motivations. As discussed in section 2.2.2, BGMRES-R relies on the subspace decomposition $\mathcal{V}_1 = \hat{\mathcal{V}}_1$ (relation (2.8)). At the first iteration of each cycle, $k_1 = p$ directions are effectively considered in the block orthonormalization procedure, including preconditioning and matrix-vector product phases. In BGMRES-R the norm minimization property induces a nonincreasing behavior of the number of selected directions k_j in a given cycle, as shown later in Proposition 3.3. However, performing no deflation at restart ($k_1 = p, d_1 = 0$) leads to a nonmonotone behavior of k_j along cycles (see the top-right panel of Figure 5.1 for an illustration), which may induce a significant additional computational overhead if the method is often restarted. The situation with possibly multiple cycles is precisely of interest in real life applications since a moderate restart size m is usually selected to limit the memory requirements when large-scale problems are considered and/or when the number of right-hand sides p is large. To circumvent this difficulty, we propose to incorporate the subspace decomposition at the beginning of each cycle of the block Krylov subspace method, leading to

$$\begin{bmatrix}\mathcal{V}_1 & P_0\end{bmatrix} = \hat{\mathcal{V}}_1 \mathcal{F}_1, \tag{3.1}$$

with $k_1 + d_1 = p$, $\mathcal{V}_1 \in \mathbb{C}^{n\times k_1}$, $P_0 \in \mathbb{C}^{n\times d_1}$, $\mathcal{F}_1 \in \mathbb{C}^{p\times p}$ with $d_1 \neq 0$ in general. The purpose of this whole section is to analyze the properties of the resulting modified block flexible GMRES with deflation at each iteration. First, we will show in section 3.4 that performing this subspace decomposition at the beginning of each cycle will ensure a nonincreasing behavior for k_j, the number of selected directions along cycles, which is a desirable property. This is a major difference between BFGMRES-S and BGMRES-R. Second, it turns out that this modification allows us to easily incorporate truncation in the block Krylov subspace method, as shown later in section 3.6. This is particularly useful when the number of right-hand sides is large. Third, we extend the block Krylov subspace method to the case of variable preconditioning, a mandatory feature when, e.g., iterative methods are used as preconditioners, as investigated later in section 5. This last property is described next.

3.2. Flexible deflated Arnoldi relation. In a given cycle of the modified block Krylov subspace method, we assume that the preconditioning operation at iteration j ($1 \leq j \leq m$) can be represented as $Z_j = M_j^{-1} V_j$, where $Z_j \in \mathbb{C}^{n\times k_j}$, $V_j \in \mathbb{C}^{n\times k_j}$, and $M_j \in \mathbb{C}^{n\times n}$ is supposed to be nonsingular. In this setting, the block orthonormalization procedure then leads to the following relation:

$$A\mathcal{Z}_j = \hat{\mathcal{V}}_{j+1} \hat{\mathcal{H}}_j, \tag{3.2}$$

where $\mathcal{Z}_j \in \mathbb{C}^{n\times s_j}$ (see Algorithm 1 for further details). Equation (3.2)—later called the flexible deflated Arnoldi relation—can be stated as

$$A\mathcal{Z}_j = \begin{bmatrix}\mathcal{V}_{j+1} & P_j\end{bmatrix} \mathcal{H}_j,$$

[3]The suffix "S" is used to emphasize that the method is based on a subspace selection at each iteration, in both the standard and truncated cases.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

where $\begin{bmatrix}\mathcal{V}_{j+1} & P_j\end{bmatrix}$ is defined as in (2.7) and $\mathcal{H}_j = \mathcal{F}_{j+1}^H \hat{\mathcal{H}}_j$. Based on this flexible deflated Arnoldi relation, the block Krylov subspace method will minimize $||B-AX||_F$ over the space $X_0 + \text{range}(\mathcal{Z}_j Y D)$ with $Y \in \mathbb{C}^{s_j \times p}$.

3.3. Representation of the scaled initial block residual. At iteration j of a given cycle of BFGMRES-S ($1 \leq j \leq m$), we recursively define the quantity $\hat{\Lambda}_j \in \mathbb{C}^{(s_j+p)\times p}$ as

$$\hat{\Lambda}_j = \begin{bmatrix} \mathcal{F}_j^H \hat{\Lambda}_{j-1} \\ 0_{k_j \times p} \end{bmatrix}. \tag{3.3}$$

In the next lemma we derive the representation of the scaled initial block residual $R_0 D^{-1}$ with respect to the $\hat{\mathcal{V}}_{j+1}$ basis.

LEMMA 3.1. *In the modified block flexible GMRES with deflation at each iteration (BFGMRES-S), the scaled initial block residual $R_0 D^{-1}$ can be expressed in the $\hat{\mathcal{V}}_{j+1}$ basis as*

$$R_0 D^{-1} = \hat{\mathcal{V}}_{j+1} \hat{\Lambda}_j, \tag{3.4}$$

with $\hat{\Lambda}_j$ defined as in (3.3).

Proof. We prove this lemma by induction. Let $\mathcal{A}_j$ denote the assumption $R_0 D^{-1} = \hat{\mathcal{V}}_{j+1} \hat{\Lambda}_j$ at index j. We note that $\mathcal{A}_0$ holds by construction (see relation (2.1)). We suppose that $\mathcal{A}_{j-1}$ is satisfied and want to prove that $\mathcal{A}_{j-1}$ implies $\mathcal{A}_j$. Due to (2.7) and the unitary character of $\mathcal{F}_j$, the quantity $\hat{\mathcal{V}}_j \hat{\Lambda}_{j-1}$ can be expressed as

$$\hat{\mathcal{V}}_j \hat{\Lambda}_{j-1} = \begin{bmatrix}\mathcal{V}_j & P_{j-1}\end{bmatrix} \mathcal{F}_j^H \hat{\Lambda}_{j-1},$$

which can be written as

$$\begin{aligned} \hat{\mathcal{V}}_j \hat{\Lambda}_{j-1} &= \begin{bmatrix}\mathcal{V}_j & P_{j-1} & \hat{V}_{j+1}\end{bmatrix} \begin{bmatrix} \mathcal{F}_j^H \hat{\Lambda}_{j-1} \\ 0_{k_j \times p} \end{bmatrix} \\ &= \hat{\mathcal{V}}_{j+1} \hat{\Lambda}_j, \end{aligned}$$

due to (2.5) and (3.3), respectively. Since $\hat{\mathcal{V}}_j \hat{\Lambda}_{j-1} = R_0 D^{-1}$, $\mathcal{A}_j$ is then satisfied. ∎

Due to the initial subspace decomposition (3.1), we remark that the representation of the scaled initial block residual in the $\hat{\mathcal{V}}_{j+1}$ basis in BFGMRES-S involves the matrices $\mathcal{F}_l$ ($1 \leq l \leq j$). In BGMRES-R this representation differs (compare relations (2.9) and (3.3), respectively).

3.4. Minimization property. We denote by $Y_j \in \mathbb{C}^{s_j \times p}$ the solution of the reduced minimization problem $\mathcal{P}_s$ considered in BFGMRES-S:

$$\mathcal{P}_s : Y_j = \underset{Y \in \mathbb{C}^{s_j \times p}}{\text{argmin}} \, ||\hat{\Lambda}_j - \hat{\mathcal{H}}_j Y||_F, \tag{3.5}$$

with $\hat{\mathcal{H}}_j$ and $\hat{\Lambda}_j$ defined in (3.2) and (3.3), respectively. We denote by $\hat{\mathcal{R}}_j \in \mathbb{C}^{(s_j+p)\times p}$ the block residual of the reduced least-squares problem $\mathcal{P}_s$, i.e., $\hat{\mathcal{R}}_j = \hat{\Lambda}_j - \hat{\mathcal{H}}_j Y_j$ ($1 \leq j \leq m$), and define $\hat{\mathcal{R}}_0 \in \mathbb{C}^{p\times p}$ as $\hat{\mathcal{R}}_0 = \hat{\Lambda}_0$. We analyze in Proposition 3.2 the norm minimization property occurring in BFGMRES-S.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

PROPOSITION 3.2. *In the modified version of the block Krylov subspace method with deflation at each iteration (BFGMRES-S), solving the reduced minimization problem $\mathcal{P}_s$ of* (3.5) *amounts to minimizing the Frobenius norm of the block true residual $||B - AX||_F$ over the space $X_0 + \text{range}(\mathcal{Z}_j Y D)$ at iteration j ($1 \leq j \leq m$) of a given cycle, i.e.,*

$$\underset{Y \in \mathbb{C}^{s_j \times p}}{\text{argmin}} ||\hat{\Lambda}_j - \hat{\mathcal{H}}_j Y||_F = \underset{Y \in \mathbb{C}^{s_j \times p}}{\text{argmin}} ||R_0 D^{-1} - A\mathcal{Z}_j Y||_F \tag{3.6}$$

$$= \underset{Y \in \mathbb{C}^{s_j \times p}}{\text{argmin}} ||B - A(X_0 + \mathcal{Z}_j Y D)||_F, \tag{3.7}$$

with $\hat{\mathcal{H}}_j$ and $\hat{\Lambda}_j$ defined in (3.2) *and* (3.3) *respectively.*

Proof. The proof follows the same lines as that of Proposition 2.1, now using relation (3.4) and the flexible deflated Arnoldi relation (3.2). ∎

3.5. Behavior of the number of selected k_j directions along convergence. We prove the important property that the number of new directions to consider in BFGMRES-S enjoys a nonincreasing behavior along convergence, as stated in Proposition 3.3.

PROPOSITION 3.3. *Denote by $k_{j,c}$ the number of Krylov directions effectively considered as best directions to keep at the jth iteration of the cth cycle of BFGMRES-S ($1 \leq j \leq m$ and $c \geq 1$), and assume that $\mathcal{Z}_j$ is of full column rank at iteration j of cycle c. Then the following relations are satisfied:*

$$\forall c, \quad k_{j+1,c} \leq k_{j,c}, \tag{3.8}$$

$$\forall c, \quad k_{1,c+1} = k_{m+1,c}. \tag{3.9}$$

Proof. BFGMRES-S is based on a standard norm minimization procedure, as recalled in Proposition 3.2. Hence at iteration j of cycle c, $R_j D^{-1}$ can be expressed as

$$R_j D^{-1} = (I_n - \mathcal{W}_j \mathcal{W}_j^H) R_{j-1} D^{-1},$$

where $\mathcal{W}_j \in \mathbb{C}^{n \times s_j}$ denotes a matrix whose columns form an orthonormal basis of $\text{range}(A\mathcal{Z}_j)$; see, e.g., [17, section 3.1]. From [28, Theorem 3.3.16] we conclude that the singular values of the scaled block true residual are monotonically decreasing; i.e.,

$$\forall\, i \mid 1 \leq i \leq p, \quad \sigma_i(R_j D^{-1}) \leq \sigma_i(R_{j-1} D^{-1}). \tag{3.10}$$

As stated in section 2.2.4 (relation (2.15)), the determination of $k_{j+1,c}$ is directly related to the singular values of $R_j D^{-1}$ in the cycle c. Hence from the inequality (3.10) we immediately deduce the relation (3.8). Finally the equality (3.9) is just due to the initial subspace decomposition (3.1) performed at the beginning of the $(c+1)$th cycle in BFGMRES-S. ∎

We deduce from Proposition 3.3 that we ensure a monotonically nonincreasing behavior for the number of k_j selected directions *along convergence* (as depicted later in the bottom-left panel of Figure 5.1) in BFGMRES-S. This is a major difference from BGMRES-R, where a nonincreasing behavior of k_j is guaranteed only inside a cycle and not along cycles. Indeed the equality (3.9) is not satisfied in BGMRES-R due to the initial subspace decomposition (2.8). Hence BFGMRES-S is not equivalent to BGMRES-R if deflation at the beginning of a cycle occurs.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

3.6. Incorporating truncation. We first detail the subspace selection in BFGMRES-S when truncation in operations is performed, and then discuss consequences for the convergence properties. Truncation in BFGMRES-S corresponds to imposing an upper bound on the number of directions that we keep in the set of active directions. This constraint is imposed both in the initial subspace decomposition ($k_1 \leq p_f$, where $1 \leq p_f \leq p$) and at each iteration of the current cycle ($k_{j+1} \leq p_f$, $1 \leq j \leq m$). This mainly aims to reduce the computational cost of the cycle. Truncation implies just a modified selection of k_{j+1} and d_{j+1}, whereas $\mathcal{F}_{j+1}$ is obtained similarly as in section 2.2.4. More precisely, using the notation of section 2.2.4, we first choose the relative deflation threshold ε_d and define $p_d \in \mathbb{N}$ according to

$$\sigma_l(\hat{\mathcal{R}}_j) > \varepsilon_d \; tol \quad \forall \, l \, \text{such that} \, 1 \leq l \leq p_d. \tag{3.11}$$

Truncation then consists of defining k_{j+1} as $k_{j+1} = \min(p_d, p_f)$ and setting d_{j+1} as $d_{j+1} = p - k_{j+1}$. When $p_d > p_f$ we note that the inequality $\sigma_l(\hat{\mathcal{R}}_j) \leq \varepsilon_d \; tol$ does not hold for $p_f < l \leq p_d$. Hence the combination of residuals that have not approximately converged are indeed deflated. As in the nontruncated case, the corresponding directions are kept and later introduced if needed. We remark that both Propositions 3.2 and 3.3 hold in the truncated case (see the bottom-right panel of Figure 5.1 for an illustration). We stress the fact that no directions are discarded; this is the major difference with BFGMREST(m), a flexible variant of BFGMRES(m) based on deflation and truncation performed at restart only [13, section 3.2.1 and Algorithm 4].[4] Nevertheless, due to truncation, BFGMRES-S may require more iterations to converge than does its nontruncated version. However, this drawback has to be weighed against the reduced computational cost of the iterations when $p_d > p_f$. The subspace selection based on truncation is summarized later in section 4, Algorithm 2. Finally, we remark that performing truncation along cycles is made possible only because of the initial subspace decomposition (3.1) realized at the beginning of each cycle in BFGMRES-S.

4. Algorithmic details, computational cost, and memory requirements. We next present the algorithmic details of the methods introduced so far in sections 2 and 3. We conclude this section by analyzing the computational cost and memory requirements of BFGMRES-S.

4.1. Deflated block Arnoldi. Algorithm 1 introduces the jth iteration of the deflated block Arnoldi procedure with block modified Gram–Schmidt, assuming that deflation has occurred at the previous iteration ($d_j \neq 0$). If not, this algorithm then reduces to the standard flexible block Arnoldi procedure that is described in, e.g., [13, Algorithm 1]. As in standard block Arnoldi, Algorithm 1 proceeds by orthonormalizing AZ_j against all the previous preconditioned Krylov directions, but additionally, orthonormalization against P_{j-1} is performed (lines 10 and 11 of Algorithm 1). The block modified Gram–Schmidt version is presented in Algorithm 1, but a version of block Arnoldi due to Ruhe [40] or block Householder orthonormalization [3, 45] could be used as well.

4.2. Subspace decomposition. The subspace decomposition at the heart of the deflation at each iteration is described in Algorithm 2 and includes the possibility

[4] In addition, we note that BFGMRES-S can use truncation at each iteration, whereas BFGMREST(m) can use truncation only at the beginning of each cycle.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Algorithm 1. jth iteration of flexible deflated block Arnoldi with block modified Gram–Schmidt: Computation of $\hat{\mathcal{V}}_{j+1}$, $\mathcal{Z}_j$, and $s_j \in \mathbb{N}$ with $V_i \in \mathbb{C}^{n\times k_i}$ such that $V_i^H V_i = I_{k_i}$ $(1 \le i \le j)$, $p = k_j + d_j$, $P_{j-1} \in \mathbb{C}^{n\times d_j}$, and $[V_1, \ldots, V_j, P_{j-1}]^H [V_1, \ldots, V_j, P_{j-1}] = I_{s_{j-1}+p}$.

Define $s_{j-1} = \sum_{l=1}^{j-1} k_l$ $(s_0 = 0)$.
Choose preconditioning operator M_j^{-1}.
$Z_j = M_j^{-1} V_j$
$S = A Z_j$
# Orthogonalization of S with respect to $[V_1, \ldots, V_j, P_{j-1}]$
for $i = 1, \ldots, j$ **do**
  $H_{i,j} = V_i^H S$
  $S = S - V_i H_{i,j}$
end for
$H_p = P_{j-1}^H S$
$S = S - P_{j-1} H_p$
Define $H_j \in \mathbb{C}^{(s_{j-1}+p)\times k_j}$ as $H_j^T = [H_{1,j}, \ldots, H_{j,j}, H_p]^T$.
Compute the QR decomposition of S as $S = QT$, $Q \in \mathbb{C}^{n\times k_j}$, and $T \in \mathbb{C}^{k_j \times k_j}$.
Set $\hat{V}_{j+1} = Q$, $H_{j+1,j} = T$.
Define $s_j = s_{j-1} + k_j$.
Define $\mathcal{Z}_j \in \mathbb{C}^{n\times s_j}$ as $\mathcal{Z}_j = [Z_1, \ldots, Z_j]$, $\mathcal{V}_j \in \mathbb{C}^{n\times s_j}$ as $\mathcal{V}_j = [V_1, \ldots, V_j]$, and $\hat{\mathcal{V}}_{j+1} \in \mathbb{C}^{n\times(s_j+p)}$ as $\hat{\mathcal{V}}_{j+1} = \begin{bmatrix} \mathcal{V}_j & P_{j-1} & \hat{V}_{j+1} \end{bmatrix}$ such that $A Z_j = \hat{\mathcal{V}}_{j+1} \begin{bmatrix} H_j \\ H_{j+1,j} \end{bmatrix}$.

Algorithm 2. Determination of k_{j+1}, d_{j+1}, and $\mathcal{F}_{j+1}$ $(0 \le j \le m)$.

Choose a relative deflation threshold ε_d and the upper bound p_f $(1 \le p_f \le p)$.
Compute the SVD of $\hat{\mathcal{R}}_j$ as $\hat{\mathcal{R}}_j = U \Sigma W^H$ with $U \in \mathbb{C}^{(s_j+p)\times p}$, $\Sigma \in \mathbb{C}^{p\times p}$, and $W \in \mathbb{C}^{p\times p}$.
Select p_d singular values of $\hat{\mathcal{R}}_j$ such that $\sigma_l(\hat{\mathcal{R}}_j) > \varepsilon_d \, tol$ for all l such that $1 \le l \le p_d$.
Set $k_{j+1} = \min(p_d, p_f)$ and $d_{j+1} = p - k_{j+1}$.
Define $U_p \in \mathbb{C}^{p\times p}$ as $U_p = U(s_j + 1 : s_j + p, 1 : p)$.
Compute the QR decomposition of U_p as $U_p = F_j T_j$, with $F_j \in \mathbb{C}^{p\times p}$, $F_j^H F_j = I_p$.
Define $\mathcal{F}_{j+1} \in \mathbb{C}^{(s_j+p)\times(s_j+p)}$ as $\mathcal{F}_{j+1} = \begin{bmatrix} I_{s_j} & 0_{s_j\times p} \\ 0_{p\times s_j} & F_j \end{bmatrix}$.

of truncation. The deflation threshold ε_d is usually fixed and does not depend on the cycle. The nontruncated variant of the algorithm introduced in section 2.2.4 is simply recovered by setting $p_f = p$. In practice, we point out that only the $p \times p$ F_j matrix has to be stored in memory.

4.3. Algorithm of modified block flexible GMRES with deflation at each iteration. Algorithm 3 introduces the modified block flexible GMRES method with deflation at each iteration. This algorithm is later named BFGMRES-S(m, p_f), where m denotes the maximal number of iterations performed in a given cycle and p_f the upper bound on the number of directions to consider at iteration j of a given cycle when performing truncation $(1 \le p_f \le p)$. The nontruncated variant is sim-

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Algorithm 3. BFGMRES-S(m, p_f).

1: Choose a convergence threshold tol, a relative deflation threshold ε_d, the size of the restart m, the maximum number of cycles $cycle_{\max}$, and maximal number of directions to keep p_f.
2: Choose an initial guess $X_0 \in \mathbb{C}^{n\times p}$.
3: Compute the initial block residual $R_0 = B - AX_0$.
4: Define the scaling diagonal matrix $D \in \mathbb{C}^{p\times p}$ as $D = \mathrm{diag}(b_1, \ldots, b_p)$ with $b_l = ||B(:, l)||_2$ for l such that $1 \le l \le p$.
5: Set $s_0 = 0$.
6: **for** $cycle = 1, cycle_{\max}$ **do**
7: Compute the QR decomposition of $R_0 D^{-1}$ as $R_0 D^{-1} = \hat{\mathcal{V}}_1 \hat{\Lambda}_0$ with $\hat{\mathcal{V}}_1 \in \mathbb{C}^{n\times p}$ and $\hat{\Lambda}_0 \in \mathbb{C}^{p\times p}$.
8: Determine deflation unitary matrix $\mathcal{F}_1 \in \mathbb{C}^{p\times p}$ and k_1, d_1 such that $k_1 + d_1 = p$ (see Algorithm 2), and set $s_1 = k_1$.
9: Define $\begin{bmatrix}\mathcal{V}_1 & P_0\end{bmatrix} = \hat{\mathcal{V}}_1 \mathcal{F}_1$, with $\mathcal{V}_1 \in \mathbb{C}^{n\times s_1}$ ($P_0 \in \mathbb{C}^{n\times d_1}$) as the first s_1 (last d_1) columns of $\hat{\mathcal{V}}_1 \mathcal{F}_1$, and define $V_1 = \mathcal{V}_1$.
10: **for** $j = 1, m$ **do**
11: *Completion of $\hat{\mathcal{V}}_{j+1}$, $\mathcal{Z}_j$, and $\hat{\mathcal{H}}_j$:* Apply Algorithm 1 to obtain $\mathcal{Z}_j \in \mathbb{C}^{n\times s_j}$, $\hat{\mathcal{V}}_{j+1} \in \mathbb{C}^{n\times(s_j+p)}$, and $\hat{\mathcal{H}}_j \in \mathbb{C}^{(s_j+p)\times s_j}$ such that

$$A\mathcal{Z}_j = \hat{\mathcal{V}}_{j+1}\hat{\mathcal{H}}_j \quad \text{with} \quad \hat{\mathcal{V}}_{j+1} = \left[V_1, V_2, \ldots, V_j, P_{j-1}, \hat{V}_{j+1}\right].$$

12: Set $\hat{\Lambda}_j \in \mathbb{C}^{(s_j+p)\times p}$ as $\hat{\Lambda}_j = \begin{bmatrix}\mathcal{F}_j^H \hat{\Lambda}_{j-1} \\ 0_{k_j\times p}\end{bmatrix}$.
13: Solve the minimization problem $\mathcal{P}_s$: $Y_j = \mathrm{argmin}_{Y\in\mathbb{C}^{s_j\times p}} ||\hat{\Lambda}_j - \hat{\mathcal{H}}_j Y||_F$.
14: Compute $\hat{\mathcal{R}}_j = \hat{\Lambda}_j - \hat{\mathcal{H}}_j Y_j$.
15: **if** $||\hat{\mathcal{R}}_j(:, l)||_2 \le tol,\ \forall\ l \mid 1 \le l \le p$, **then**
16: Compute $X_j = X_0 + \mathcal{Z}_j Y_j D$; stop;
17: **end if**
18: Determine deflation unitary matrix $\mathcal{F}_{j+1} \in \mathbb{C}^{(s_j+p)\times(s_j+p)}$ and k_{j+1}, d_{j+1} such that $k_{j+1} + d_{j+1} = p$ (see Algorithm 2).
19: Set $s_{j+1} = s_j + k_{j+1}$.
20: Define $\begin{bmatrix}\mathcal{V}_{j+1} & P_j\end{bmatrix} = \hat{\mathcal{V}}_{j+1}\mathcal{F}_{j+1}$, with $\mathcal{V}_{j+1} \in \mathbb{C}^{n\times s_{j+1}}$ (or $P_j \in \mathbb{C}^{n\times d_{j+1}}$) as the first s_{j+1} (or last d_{j+1}) columns of $\hat{\mathcal{V}}_{j+1}\mathcal{F}_{j+1}$.
21: Define $\mathcal{H}_j = \mathcal{F}_{j+1}^H \hat{\mathcal{H}}_j$, with $\mathcal{H}_j \in \mathbb{C}^{(s_j+p)\times s_j}$.
22: **end for**
23: $X_m = X_0 + \mathcal{Z}_m Y_m D$
24: $R_m = B - AX_m$
25: Set $R_0 = R_m$ and $X_0 = X_m$.
26: **end for**

ply recovered if $p_f = p$ is satisfied. In such a case, the algorithm is simply named BFGMRES-S(m).

A comparison of BFGMRES-R due to Robbé and Sadkane [39] (Algorithm 4, given in the appendix for convenience) with BFGMRES-S (Algorithm 3) reveals the three main differences discussed in section 3: the initial subspace decomposition (performed at lines 8 and 9), the modified representation of the reduced right-

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

hand side (line 12), and the resulting different minimization problem to be solved (line 13).

4.4. Computational cost and memory requirements. The question of the total computational cost of BFGMRES-S is now addressed. For that purpose we summarize in Table 4.1 the costs occurring during a given cycle of BFGMRES-S(m, p_f) (considering Algorithms 1, 2, and 3), excluding matrix-vector products and preconditioning operations which are problem-dependent. We have included the costs proportional to both the size of the original problem n and the number of right-hand sides p, assuming a QR-factorization based on modified Gram–Schmidt and a Golub–Reinsch SVD;[5] see, e.g, [23, section 5.4.5] and [27, Appendix C] for further details on operation counts. The total cost of a given cycle is then found to grow as $C_1np^2 + C_2p^3 + C_3np$, and we note that this cost is always nonincreasing along convergence due to Proposition 3.3.

Compared to BGMRES-R, additional operations are related to the computations of $\mathcal{F}_1$ and $\hat{\Lambda}_j$, operations that behave as p^3. The computation of $\begin{bmatrix}\mathcal{V}_{j+1} & P_j\end{bmatrix}$ is in practice the most expensive one in a given iteration of BFGMRES-S(m, p_f). Concerning the truncated variant, the computational cost of a cycle will be reduced only if $p_d > p_f$, since the upper bound on k_{j+1} will then be active. This situation occurs at the beginning of the convergence due to the nonincreasing behavior of the singular values of $\hat{\mathcal{R}}_j$ shown in Proposition 3.3.

TABLE 4.1

Computational cost of a cycle of BFGMRES-S(m, p_f) (Algorithm 3). This excludes the cost of matrix-vector operations and preconditioning operations.

Step	Computational cost
Computation of R_0D^{-1}	np
QR factorization of R_0D^{-1}	$2np^2 + np$
Computation of $\mathcal{F}_1$	$14p^3$
Computation of $\begin{bmatrix}\mathcal{V}_1 & P_0\end{bmatrix}$	$2np^2$
Block Arnoldi procedure[6]	C_j
Computation of $\hat{\Lambda}_j$	$2(s_{j-1} + p)^2p$
Computation of Y_j	$2s_j^3 + 3ps_j^2$
Computation of $\hat{\mathcal{R}}_j$	$(2s_j + 1)(s_j + p)p$
Computation of $\mathcal{F}_{j+1}$	$4s_jp^2 + 14p^3$
Computation of $\begin{bmatrix}\mathcal{V}_{j+1} & P_j\end{bmatrix}$	$2np^2$
Computation of $\mathcal{H}_j$	$2p^3$
Computation of X_m	$np + (2n + 1)s_mp$

Concerning storage proportional to the problem size n, BFGMRES-S(m, p_f) requires R_m, X_0, X_m, $\mathcal{V}_{m+1}$, and $\mathcal{Z}_m$ leading to a memory requirement of $2ns_m + 4np$ at the end of a given cycle. Since s_m varies from cycle to cycle, an upper bound of the memory requirement can be given as $n(2m + 1)p + 3np$ when p linear systems have to be considered at the beginning of a given cycle. We note that the storage is monotonically decreasing along convergence, a feature than can be, for instance, exploited if dynamic memory allocation is used.

[5]The Golub–Reinsch SVD decomposition $R = U\Sigma V^H$ with $R \in \mathbb{C}^{m\times n}$ requires $4mn^2 + 8n^3$ operations when only Σ and V have to be computed.

[6]Algorithm 1: The block Arnoldi method based on modified Gram–Schmidt requires $\sum_{j=1}^{m}\sum_{i=1}^{j}(4nk_ik_j + nk_j + 4nd_jk_j)$ operations (lines 6 to 11) plus $\sum_{j=1}^{m} 2nk_j^2$ operations for the QR decomposition of S (line 13). Thus $C_j = \sum_{j=1}^{m}(\sum_{i=1}^{j}(4nk_ik_j + nk_j + 4nd_jk_j) + 2nk_j^2)$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

5. Numerical experiments. We investigate the numerical behavior of block flexible Krylov subspace methods including deflation at each iteration on a challenging application in geophysics where the situation of multiple right-hand sides is common. The source terms correspond to Dirac sources in this example. Thus the block right-hand side $B \in \mathbb{C}^{n\times p}$ is extremely sparse (only one nonzero element per column), and the initial block residual corresponds to a full rank matrix. We compare both BFGMRES-R(m) and BFGMRES-S(m) with various preconditioned iterative methods based on flexible (block) GMRES(m) with a zero initial guess (X_0) and a moderate value of the restart parameter m. The iterative procedures are stopped when the following condition is satisfied:

$$\frac{||B(:,l) - AX(:,l)||_2}{||B(:,l)||_2} \leq tol \qquad \forall\ l = 1, \ldots, p.$$

A primary concern will be to evaluate whether BFGMRES-S(m) can be efficient when solving problems with multiple right-hand sides both in terms of preconditioner applications and total computational cost. Finally, the tolerance is set to $tol = 10^{-5}$ in the numerical experiments, and we fix the parameter ϵ_d of Algorithm 2 to 1.

5.1. Acoustic full waveform inversion. We focus on a specific application in geophysics related to the simulation of wave propagation phenomena on Earth [47]. Given a three-dimensional physical domain Ω_p, the propagation of a wave field in a heterogeneous medium can be modeled by the Helmholtz equation written in the frequency domain:

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} - \frac{\partial^2 u}{\partial z^2} - \frac{(2\pi f)^2}{c^2(x,y,z)} u = g_s(\mathbf{x}), \quad \mathbf{x} = (x,y,z) \in \Omega_p. \tag{5.1}$$

u represents the pressure field in the frequency domain, c the variable acoustic-wave velocity in ms^{-1}, and f the frequency in Hertz. The source term $g_s(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x_s})$ represents a harmonic point source located at (x_s, y_s, z_s). A popular approach—the perfectly matched layer formulation (PML) [8, 9]—has been used in order to obtain a satisfactory near-boundary solution, without many artificial reflections. As in [13], we consider a second-order finite difference discretization of the Helmholtz equation (5.1) on a uniform equidistant Cartesian grid of size $n_x \times n_y \times n_z$. The same stability condition (12 points per wavelength) relating f, the frequency, to h, the mesh grid size, and $c(x,y,z)$, the heterogeneous velocity field, has been considered ($12fh = \min_{(x,y,z)\in\Omega_h} c(x,y,z)$). In consequence, A is a sparse complex matrix which is non-Hermitian and nonsymmetric due to the PML formulation that leads to complex-valued variable coefficients in the partial differential equation [37, Appendix A]. The resulting linear systems are known to be challenging for iterative methods [19, 20]. We consider the same approximate geometric two-level preconditioner presented in [13], which has been shown to be relatively efficient for the solution of three-dimensional heterogeneous Helmholtz problems in geophysics. We refer the reader to [13, Algorithm 5] for a complete description of the geometric preconditioner, and to [37] for additional theoretical properties in relation to Krylov subspace methods. In this section we consider this variable two-grid preconditioner in the multiple right-hand-side case and next investigate the performance of the block flexible Krylov methods on this challenging real-life application. The numerical results have been obtained on Babel, a Blue Gene/P computer located at IDRIS (PowerPC 450, 850 MHz, with 512 MB of memory on each core), using a Fortran 90 implementation with MPI

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

in single precision arithmetic. This code was compiled by the IBM compiler suite with standard compiling options and linked with the vendor BLAS and LAPACK subroutines.

As in [13], we consider the velocity field issued from the public domain SEG/EAGE Overthrust model [2] and analyze the performance of the numerical methods at a given frequency $f = 3.64$ Hz. Both the problem dimension (about 23 million unknowns) and the maximal number of right-hand sides to be considered (128) correspond to a task that geophysicists typically must face on a daily basis. Thus efficient numerical methods must be developed for that purpose. In [13] we have considered block flexible Krylov subspace methods including deflation at restart only for this application with a reduced number of right-hand sides (from 4 to 16). We continue this detailed analysis and investigate the performance of both BFGMRES-S(m, p_f) and BFGMRES-R(m) with a larger number of right-hand sides. In addition, we consider the standard block flexible GMRES method (BFGMRES(m)), the block flexible GMRES(m) with deflation performed at restart only (BFGMRESD(m) [13, Algorithm 3]), and the block flexible GMRES(m) with deflation and truncation performed at restart only (BFGMREST(m, p_f) [13, Algorithm 4]). We also investigate a combination of BFGMRES-S and BFGMRESD. This method, later named Combined(m, p_s), corresponds to BFGMRES-S(m) at the beginning of the convergence history. Then as soon as the number of Krylov directions effectively considered at iteration j (k_j) reaches a given prescribed value (p_s), the method switches to BFGMRESD(m) at the next restart. This mainly aims at reducing the computational cost in the next cycles by performing deflation *only* at the restart instead of at each iteration. Finally the number of cores is set to $8p$, ranging from 32 for $p = 4$ to 1024 for $p = 128$. This aims at imposing the same memory constraint on each core for all numerical experiments, as in [13]. The maximal memory requested is about 488 Gb for $p = 128$.

Table 5.1 collects, in addition to iterations (It)[7] and preconditioner applications on a single vector (Pr),[8] the computational times in seconds (T). Among the different strategies, BFGMRES-S(5) most often delivers the minimal number of preconditioner applications and computational times (see italic and bold values, respectively, in Table 5.1). This clearly highlights the value of performing deflation at each iteration, both in terms of preconditioner applications and computational operations on this given application. The improvement over BFGMRES-R(5) ranges from 10% for $p = 4$ to 35% for $p = 128$, which is very satisfactory behavior. BFGMRES-S(5) is also found to be competitive with respect to methods incorporating deflation at restart only (a gain of up to 15% in terms of computational time is obtained, for instance, for $p = 8$) as well as BFGMRES-S(5,p/2) (maximal gain of 21% (for $p = 32$) when compared to BFGMREST(5,p/2)). This is a satisfactory improvement, since methods including deflation at restart only are already quite efficient in this application, as shown in [13]. We also note that the improvement over the classical block flexible GMRES method is quite large as expected (a maximal gain of about 60% is obtained for $p = 64$).

We have also considered the solution of the p linear systems given now in sequence with the FGMRES Krylov subspace method [41]. In Table 5.1, FGMRES($5p$) consists of solving the p linear systems in sequence (starting with a zero initial guess),

[7] A complete cycle of BFGMRES(m), BFGMRES-R(m), or BFGMRES-S(m) always corresponds to m iterations, whereas a complete cycle of FGMRES(mp) involves mp iterations.

[8] A complete cycle of BFGMRES(m) corresponds to mp preconditioner applications, whereas a complete cycle of either BFGMRES-R(m) or BFGMRES-S(m) corresponds to $\sum_{j=1}^{m} k_{j,c}$ preconditioner applications. A complete cycle of FGMRES(mp) requires mp preconditioner applications.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

TABLE 5.1

Acoustic full waveform inversion (SEG/EAGE Overthrust model). Case of $f = 3.64$ Hz ($h = 50$ m), with $p = 4$ to $p = 128$ right-hand sides given at once. It denotes the number of iterations, Pr the number of preconditioner applications on a single vector, and T the total computational time in seconds. The number of cores is set to 8p.

Acoustic full waveform inversion - *Grid* : $433 \times 433 \times 126$									
	$p = 4$			$p = 8$			$p = 16$		
Method	It	Pr	T	It	Pr	T	It	Pr	T
FGMRES(5p)	56	56	624	112	112	629	224	224	665
BFGMRES(5)	14	56	622	14	112	631	14	224	668
BFGMRESD(5)	14	43	489	15	70	401	15	120	371
BFGMRES-R(5)	16	44	503	16	74	431	16	134	417
BFGMRES-S(5)	16	*39*	**452**	16	*57*	**339**	18	*102*	328
BFGMREST(5,p/2)	24	48	542	23	80	447	20	140	410
BFGMRES-S(5,p/2)	16	40	459	15	68	392	17	124	384
Combined(5,p/2)	15	41	471	15	62	359	15	103	323
Combined(5,p/4)	18	41	474	15	59	346	15	*102*	**320**
	$p = 32$			$p = 64$			$p = 128$		
Method	It	Pr	T	It	Pr	T	It	Pr	T
FGMRES(5p)	434	434	670	1152	1152	925	2531	2531	1187
BFGMRES(5)	14	448	713	18	1152	962	19	2432	1187
BFGMRESD(5)	15	225	371	20	490	422	25	1015	509
BFGMRES-R(5)	18	283	466	25	618	537	28	1489	762
BFGMRES-S(5)	19	*181*	316	25	413	375	28	915	497
BFGMREST(5,p/2)	20	255	396	25	550	444	28	1125	524
BFGMRES-S(5,p/2)	16	189	310	24	444	396	29	976	523
Combined(5,p/2)	15	184	**305**	20	409	348	25	899	**442**
Combined(5,p/4)	20	191	320	20	*398*	**342**	25	*898*	448

the Euclidean norm of each residual being minimized over a subspace of maximal dimension 5p. The maximal number of iterations performed to reach the stopping criterion (5.1) on a single linear system is found to be equal to 14 (p ranging from 4 to 32), 18 ($p = 64$), and 22 ($p = 128$), respectively. These results lead to two important comments. First, whatever the number of right-hand sides considered, no restart occurs in the Krylov subspace method applied in a single right-hand side situation: FGMRES(5p) thus corresponds to a preconditioned full flexible GMRES method in such a case. This is thus ideal for FGMRES(5p), since no restart procedure that might have hampered the convergence of the method is involved. Second, we remark that the maximal number of iterations performed does depend on the number of cores. This behavior can be explained as follows. An analysis of the FGMRES Krylov subspace method with the variable two-grid preconditioner on three-dimensional heterogeneous Helmholtz problems has shown that the numerical method satisfies a strong scalability property up to a given number of cores [37]. The loss of scalability is indeed due to the symmetric Gauss-Seidel preconditioner used both in the smoother and in the approximate solution of the coarse problem. This preconditioner is based on a subdomain decoupling and thus becomes inherently less efficient when the number of cores is increasing [6]. We refer the reader to [37] and [13, section 4.2.2] for related numerical experiments and additional comments. Finally, we remark that the improvement due to block methods using deflation at each iteration over the flexible GMRES method applied on the sequence of linear systems is noticeable on this application; a maximal gain of about 62% is obtained for $p = 128$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

TABLE 5.2

Acoustic full waveform inversion (SEG/EAGE Overthrust model). Case of $f = 3.64\ Hz$ ($h = 50\ m$), with $p = 4$ to $p = 128$ right-hand sides given at once. Detailed timings (in seconds) related to orthogonalization (T_{orth}) and to preconditioning and matrix-vector products (T_{pmvp}). Here $\sigma = T_{pmvp}/T$ represents the percentage of time spent in the preconditioning and matrix-vector product phases with respect to the total computational times (T) given in Table 5.1. The number of cores is set to $8p$.

Acoustic full waveform inversion - *Grid* : $433 \times 433 \times 126$									
	$p = 4$			$p = 8$			$p = 16$		
Method	T_{orth}	T_{pmvp}	σ	T_{orth}	T_{pmvp}	σ	T_{orth}	T_{pmvp}	σ
FGMRES($5p$)	10	607	0.97	8	609	0.97	5	646	0.97
BFGMRES(5)	13	605	0.97	17	608	0.96	29	631	0.94
BFGMRESD(5)	14	470	0.96	11	384	0.96	16	348	0.94
BFGMRES-R(5)	15	480	0.95	14	408	0.95	18	386	0.93
BFGMRES-S(5)	16	428	0.95	12	317	0.94	15	299	0.91
BFGMREST(5,p/2)	16	519	0.96	10	425	0.95	11	391	0.95
BFGMRES-S(5,p/2)	15	436	0.95	10	373	0.95	15	357	0.93
Combined(5,p/2)	16	449	0.95	10	343	0.96	14	301	0.93
Combined(5,p/4)	16	449	0.95	12	328	0.95	13	298	0.93
	$p = 32$			$p = 64$			$p = 128$		
Method	T_{orth}	T_{pmvp}	σ	T_{orth}	T_{pmvp}	σ	T_{orth}	T_{pmvp}	σ
FGMRES($5p$)	9	614	0.92	15	862	0.93	25	1141	0.96
BFGMRES(5)	51	649	0.91	116	818	0.85	223	906	0.76
BFGMRESD(5)	25	334	0.90	45	354	0.84	74	385	0.76
BFGMRES-R(5)	27	417	0.89	30	428	0.80	98	565	0.74
BFGMRES-S(5)	19	275	0.87	31	300	0.80	56	348	0.70
BFGMREST(5,p/2)	16	368	0.93	29	389	0.88	50	423	0.81
BFGMRES-S(5,p/2)	17	276	0.89	31	320	0.81	58	371	0.71
Combined(5,p/2)	17	276	0.90	28	297	0.85	50	342	0.77
Combined(5,p/4)	18	286	0.89	27	288	0.84	51	341	0.76

Detailed computational timings spent in the orthogonalization phase (T_{orth}) and in both preconditioning and outer matrix-vector product phases (T_{pmvp}) are provided in Table 5.2. In addition the percentages (σ) of time spent in the preconditioning and matrix-vector product phases with respect to the total computational times are given. The analysis of σ clearly highlights that the dominant cost in all the methods is related to the preconditioning phase, which is in agreement with the main assumption of the paper. In the application, the approximate solution of the coarse linear system obtained with a symmetric Gauss–Seidel preconditioned restarted GMRES method represents the most computationally expensive part of the two-grid cycle used as a preconditioner. We refer the reader to [37] for further details on the preconditioner.

Figure 5.1 shows the evolution of k_j along convergence for the various block subspace methods in the case of $p = 32$. Regarding BFGMRESD(5) and BFGMREST(5,p/2) deflation is performed only at the beginning of each cycle; thus k_j is found to be constant in a given cycle. Variations at each iteration can happen only in BFGMRES-R(5) or in BFGMRES-S(5). As expected, BFGMRES-S(5) enjoys a nonincreasing behavior for k_j along convergence, while peaks occur for BFGMRES-R(5) at the beginning of each cycle (see Proposition 3.3). In this example the use of truncation within BFGMRES-S(5, p/2) tends to delay the beginning of the decreasing behavior of k_j. After a certain phase deflation is nevertheless active and proves to be useful.

We also remark that the use of truncation techniques in BFGMRES-S(m, p_f) leads to an efficient method. In certain cases BFGMRES-S(5, p/2) is as efficient as

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

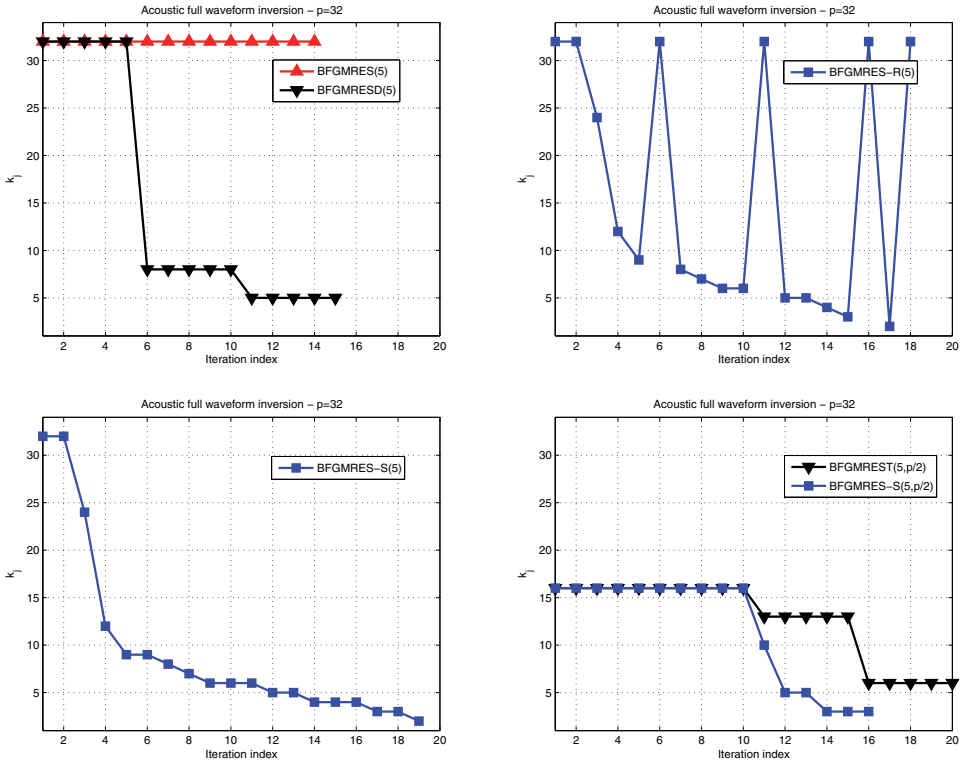


FIG. 5.1. *Acoustic full waveform inversion (SEG/EAGE Overthrust model). Case of* $p = 32$. *Evolution of* k_j *versus iterations for* $p = 32$ *in BFGMRES(5) and BFGMRESD(5) (top-left), BFGMRES-R(5) (top-right), BFGMRES-S(5) (bottom-left), and truncated variants (BFGMREST(5,p/2) and BFGMRES-S(5,p/2)) (bottom-right).*

BFGMRES-S(5) in terms of computational times (see, e.g., the case $p = 32$ in Table 5.1). This feature is really important in the given application due to the large size of the linear systems. Furthermore BFGMRES-S(5, p/2) usually requires fewer preconditioner applications than does BFGMREST(5, p/2). This satisfactory behavior has a definite reason: due to Proposition 3.2, we guarantee that the truncated variant of BFGMRES-S(m, p_f) minimizes the whole residual at each iteration (regardless of the value of p_f), whereas BFGMREST(m) chooses just a set of linear independent columns of the block residual to be minimized at each cycle. We consider that this is indeed a critical feature of the truncated variant of BFGMRES-S(m, p_f). Furthermore, as shown in Table 5.1, the Combined$(5, p_s)$ method (with $p_s = p/2$ or $p_s = p/4$) leads to further reductions in computational times and is especially appropriate when the number of right-hand sides becomes large on this given application.

Finally, in [12, section 6.1] and [30, section 3.9.3] the first five strategies (FGMRES(mp), BFGMRES(m), BFGMRESD(m), BFGMRES-R(m), and BFGMRES-S(m)) have been evaluated on an academic test case related to a two-dimensional partial differential equation (complex-valued advection-diffusion reaction problem) with a number of right-hand sides ranging from 4 to 32. A cycle of GMRES(m) has been used as a variable preconditioner in all methods. Whatever the value of the restart

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

parameter m (two values have been considered, $m = 5$ and $m = 10$), it was found that BFGMRES-S(m) always led to the minimal number of preconditioner applications and delivered the best efficiency in terms of computational operations. This is thus behavior similar to the proposed application in geophysics. We also refer the reader to [30, section 3.9] for additional numerical experiments on academic problems related to partial differential equations showing a similar trend.

6. Conclusion. We have proposed a block restarted GMRES method for the solution of non-Hermitian linear systems with multiple right-hand sides that allows both the use of deflation at each iteration and variable preconditioning. This method uses a subspace decomposition based on the singular value decomposition of the block residual of the reduced least-squares problem. This decomposition aims at selecting a set of k_j new Krylov directions at iteration j, while d_j directions are deflated (i.e., kept and reintroduced later if needed) at the same iteration. The new method ensures a nonincreasing behavior of k_j along convergence, which leads to possibly considerable computational savings with respect to the existing reference method [39]. We have also proposed a variant based on truncation. All these features are particularly of interest when tackling the solution of large-scale linear systems with many right-hand sides. BFGMRES-S has proved to be efficient in terms of both preconditioner applications and computational operations on an application related to geophysics. Often, *but not always*, it has been found superior to recent block flexible methods including deflation at restart only. We would like to emphasize that, when large restart sizes m or large numbers of right-hand sides p are considered, the cost of orthogonalization can become significant. In consequence this may potentially decrease the value of performing deflation at each iteration. Nevertheless, in this paper, satisfactory behavior has been observed on an industrial simulation, where large linear systems with multiple right-hand sides have been successfully solved in a parallel distributed memory environment. Further reductions in terms of computational times have been obtained by combining methods including deflation at each iteration and deflation at restart only in a second phase.

It is worth noting that the theoretical properties of BFGMRES-S hold for any unitary matrix $\mathcal{F}_{j+1}$. Hence different subspace decompositions could be investigated. We also note that the analysis proposed in this paper can be extended as well to other block Krylov subspace methods based on a norm minimization property, such as block FOM [38], block GCRO [49], and block simpler GMRES [32]. All these methods do rely on block orthogonalizations that require global communications. These latter operations usually become a bottleneck on massively parallel platforms, and we plan in the near future to investigate algorithmic variants, where these global communications can be overlapped with calculations or local communications. This is especially interesting for large-scale problems.

To give a broader picture of the performance of the block Krylov subspace methods investigated here, we finally mention that a comparison with flexible variants of block Lanczos algorithms including deflation at each iteration should be performed. This is the topic of a forthcoming study.

Appendix. Algorithm 4 shows the restarted block GMRES method with deflation at each iteration in the case of variable preconditioning that is considered in section 5. This algorithm is named BFGMRES-R(m). We note that the original algorithm [39, Algorithm 2] is simply recovered if each preconditioning operator M_j is chosen as the identity operator I_n in Algorithm 1.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Algorithm 4. BFGMRES-R(m) [39].

1: Choose a convergence threshold tol, a relative deflation threshold ε_d, the size of the restart m, and the maximum number of cycles $cycle_{\max}$.
2: Choose an initial guess $X_0 \in \mathbb{C}^{n\times p}$.
3: Compute the initial block residual $R_0 = B - AX_0$.
4: Define the scaling diagonal matrix $D \in \mathbb{C}^{p\times p}$ as $D = \mathrm{diag}(b_1, \ldots, b_p)$ with $b_l = ||B(:,l)||_2$ for l such that $1 \le l \le p$.
5: Set $s_0 = 0$.
6: **for** $cycle = 1, cycle_{\max}$ **do**
7: Compute the QR decomposition of $R_0 D^{-1}$ as $R_0 D^{-1} = \hat{\mathcal{V}}_1 \hat{\Lambda}_0$ with $\hat{\mathcal{V}}_1 \in \mathbb{C}^{n\times p}$ and $\hat{\Lambda}_0 \in \mathbb{C}^{p\times p}$.
8: Set $k_1 = p$, $d_1 = 0$, and $s_1 = k_1$.
9: Define[9] $\begin{bmatrix}\mathcal{V}_1 & P_0\end{bmatrix} = \hat{\mathcal{V}}_1$, with $\mathcal{V}_1 \in \mathbb{C}^{n\times s_1}$ ($P_0 \in \mathbb{C}^{n\times d_1}$) as the first s_1 (last d_1) columns of $\hat{\mathcal{V}}_1$, and define $V_1 = \mathcal{V}_1$.
10: **for** $j = 1, m$ **do**
11: *Completion of $\hat{\mathcal{V}}_{j+1}$, $\mathcal{Z}_j$, and $\hat{\mathcal{H}}_j$:* Apply Algorithm 1 to obtain $\mathcal{Z}_j \in \mathbb{C}^{n\times s_j}$, $\hat{\mathcal{V}}_{j+1} \in \mathbb{C}^{n\times(s_j+p)}$, and $\hat{\mathcal{H}}_j \in \mathbb{C}^{(s_j+p)\times s_j}$ such that

$$A\mathcal{Z}_j = \hat{\mathcal{V}}_{j+1}\hat{\mathcal{H}}_j \quad \text{with} \quad \hat{\mathcal{V}}_{j+1} = \left[V_1, V_2, \ldots, V_j, P_{j-1}, \hat{V}_{j+1}\right].$$

12: Set $\hat{\Lambda}_j \in \mathbb{C}^{(s_j+p)\times p}$ as $\hat{\Lambda}_j = \begin{bmatrix}\hat{\Lambda}_0 \\ 0_{s_j\times p}\end{bmatrix}$.
13: Solve the minimization problem $\mathcal{P}_r$: $Y_j = \mathrm{argmin}_{Y\in\mathbb{C}^{s_j\times p}} ||\hat{\Lambda}_j - \hat{\mathcal{H}}_j Y||_F$.
14: Compute $\hat{\mathcal{R}}_j = \hat{\Lambda}_j - \hat{\mathcal{H}}_j Y_j$.
15: **if** $||\hat{\mathcal{R}}_j(:,l)||_2 \le tol \ \forall\ l \mid 1 \le l \le p$, **then**
16: Compute $X_j = X_0 + \mathcal{Z}_j Y_j D$; stop;
17: **end if**
18: Determine deflation unitary matrix $\mathcal{F}_{j+1} \in \mathbb{C}^{(s_j+p)\times(s_j+p)}$ and k_{j+1}, d_{j+1} such that $k_{j+1} + d_{j+1} = p$ (see Algorithm 2 with $p_f = p$).
19: Set $s_{j+1} = s_j + k_{j+1}$.
20: Define $\begin{bmatrix}\mathcal{V}_{j+1} & P_j\end{bmatrix} = \hat{\mathcal{V}}_{j+1}\mathcal{F}_{j+1}$, with $\mathcal{V}_{j+1} \in \mathbb{C}^{n\times s_{j+1}}$ (or $P_j \in \mathbb{C}^{n\times d_{j+1}}$) as the first s_{j+1} (or last d_{j+1}) columns of $\hat{\mathcal{V}}_{j+1}\mathcal{F}_{j+1}$.
21: Define $\mathcal{H}_j = \mathcal{F}_{j+1}^H \hat{\mathcal{H}}_j$, with $\mathcal{H}_j \in \mathbb{C}^{(s_j+p)\times s_j}$.
22: **end for**
23: $X_m = X_0 + \mathcal{Z}_m Y_m D$
24: $R_m = B - AX_m$
25: Set $R_0 = R_m$ and $X_0 = X_m$.
26: **end for**

Acknowledgments. The authors would like to thank Michele Benzi and the two referees for their comments and suggestions, which considerably helped to improve the manuscript. The authors would like to acknowledge GENCI (Grand Equipement National de Calcul Intensif) for the donation of computing hours on the IBM Blue Gene/P computer at IDRIS, France. The authors were granted access to the HPC resources of IDRIS under allocation 2012065068 and 2013065068 made by GENCI.

[9] We have made the abuse of notation $\begin{bmatrix}\mathcal{V}_1 & P_0\end{bmatrix} = \hat{\mathcal{V}}_1$ to allow an easy-to-read comparison with line 9 of Algorithm 3. In BFGMRES-R(m) we have $\mathcal{V}_1 = \hat{\mathcal{V}}_1$ and $P_0 = [\,]$ in practice.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

REFERENCES

[1] J. I. Aliaga, D. L. Boley, R. W. Freund, and V. Hernández, *A Lanczos-type method for multiple starting vectors*, Math. Comput., 69 (2000), pp. 1577–1601.

[2] F. Aminzadeh, J. Brac, and T. Kunz, *3D Salt and Overthrust Models*, Society of Exploration Geophysicists, Tulsa, OK, 1997.

[3] J. Baglama, *Augmented block Householder Arnoldi method*, Linear Algebra Appl., 429 (2008), pp. 2315–2334.

[4] Z. Bai, D. Day, and Q. Ye, *ABLE: An adaptive block Lanczos for non-Hermitian eigenvalue problems*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 1060–1082.

[5] A. H. Baker, J. M. Dennis, and E. R. Jessup, *On improving linear solver performance: A block variant of GMRES*, SIAM J. Sci. Comput., 27 (2006), pp. 1608–1626.

[6] A. H. Baker, R. D. Falgout, T. V. Kolev, and U. M. Yang, *Multigrid smoothers for ultraparallel computing*, SIAM J. Sci. Comput., 33 (2011), pp. 2864–2887.

[7] G. Barbella, F. Perotti, and V. Simoncini, *Block Krylov subspace methods for the computation of structural response to turbulent wind*, Comput. Methods Appl. Mech. Engrg., 200 (2011), pp. 2067–2082.

[8] J.-P. Berenger, *A perfectly matched layer for absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[9] J.-P. Berenger, *Three-dimensional perfectly matched layer for absorption of electromagnetic waves*, J. Comput. Phys., 127 (1996), pp. 363–379.

[10] W. E. Boyse and A. A. Seidl, *A block QMR method for computing multiple simultaneous solutions to complex symmetric systems*, SIAM J. Sci. Comput., 17 (1996), pp. 263–274.

[11] P. A. Businger and G. Golub, *Linear least squares solutions by Householder transformations*, Numer. Math., 7 (1965), pp. 269–276.

[12] H. Calandra, S. Gratton, R. Lago, and X. Vasseur, *A Deflated Minimal Block Residual Method for the Solution of non-Hermitian Linear Systems with Multiple Right-Hand Sides*, Technical Report TR/PA/12/45, CERFACS, Toulouse, France, 2012.

[13] H. Calandra, S. Gratton, J. Langou, X. Pinel, and X. Vasseur, *Flexible variants of block restarted GMRES methods with application to geophysics*, SIAM J. Sci. Comput., 34 (2012), pp. A714–A736.

[14] J. Cullum and T. Zhang, *Two-sided Arnoldi and nonsymmetric Lanczos algorithms*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 303–319.

[15] R. D. da Cunha and D. Becker, *Dynamic block GMRES: An iterative method for block linear systems*, Adv. Comput. Math., 27 (2007), pp. 423–448.

[16] L. Du, T. Sogabe, B. Yu, Y. Yamamoto, and S.-L. Zhang, *A block IDR(s) method for nonsymmetric linear systems with multiple right-hand sides*, J. Comput. Appl. Math., 235 (2011), pp. 4095–4106.

[17] M. Eiermann and O. Ernst, *Geometric aspects of the theory of Krylov subspace methods*, Acta Numer., 10 (2001), pp. 251–312.

[18] H. Elman, O. Ernst, D. O'Leary, and M. Stewart, *Efficient iterative algorithms for the stochastic finite element method with application to acoustic scattering*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1037–1055.

[19] Y. A. Erlangga, *Advances in iterative methods and preconditioners for the Helmholtz equation*, Arch. Comput. Methods Engrg., 15 (2008), pp. 37–66.

[20] O. Ernst and M. J. Gander, *Why it is difficult to solve Helmholtz problems with classical iterative methods*, in Numerical Analysis of Multiscale Problems, O. Lakkis, I. Graham, T. Hou, and R. Scheichl, eds., Springer, New York, 2011, pp. 325–361.

[21] R. W. Freund, *Krylov-subspace methods for reduced-order modeling in circuit simulation*, J. Comput. Appl. Math., 123 (2000), pp. 395–421.

[22] R. W. Freund and M. Malhotra, *A block QMR algorithm for non-Hermitian linear systems with multiple right-hand sides*, Linear Algebra Appl., 254 (1997), pp. 119–157.

[23] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.

[24] A. El Guennouni, K. Jbilou, and H. Sadok, *A block version of BICGSTAB for linear systems with multiple right-hand sides*, Electron. Trans. Numer. Anal., 16 (2003), pp. 129–142.

[25] M. H. Gutknecht, *Block Krylov space methods for linear systems with multiple right-hand sides: An introduction*, in Modern Mathematical Models, Methods and Algorithms for Real World Systems, A. H. Siddiqi, I. S. Duff, and O. Christensen, eds., Anamaya Publishers, New Delhi, India, 2006, pp. 420–447.

[26] M. H. Gutknecht and T. Schmelzer, *The block grade of a block Krylov space*, Linear Algebra Appl., 430 (2009), pp. 174–185.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[27] N. J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008.
[28] R. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[29] A. Khabou, *Solveur itératif haute performance pour les systèmes linéaires avec seconds membres multiples*, Master's thesis, Department of Computer Science, University of Bordeaux I, 2009.
[30] R. Lago, *A Study on Block Flexible Iterative Solvers with Applications to Earth Imaging Problem in Geophysics*, Ph.D. thesis, CERFACS, Toulouse, France, 2013.
[31] J. Langou, *Iterative methods for solving linear systems with multiple right-hand sides*, Ph.D. thesis, Department of Mathematics, CERFACS, Toulouse, France 2003.
[32] H. Liu and B. Zhong, *A simpler block GMRES for nonsymmetric systems with multiple right-hand sides*, Electron. Trans. Numer. Anal., 30 (2008), pp. 1–9.
[33] D. Loher, *Reliable Nonsymmetric Block Lanczos Algorithms*, Ph.D. thesis, Department of Mathematics, Swiss Federal Institute of Technology Zurich (ETHZ), Zurich, Switzerland, 2006.
[34] M. Malhotra, R. W. Freund, and P. M. Pinsky, *Iterative solution of multiple radiation and scattering problems in structural acoustics using a block quasi-minimal residual algorithm*, Comput. Methods Appl. Mech. Engrg., 146 (1997), pp. 173–196.
[35] A. A. Nikishin and A. Yu. Yeremin, *Variable block CG algorithms for solving large sparse symmetric positive definite linear systems on parallel computers,* I: *General iterative scheme*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1135–1153.
[36] D. P. O'Leary, *The block conjugate gradient algorithm and related methods*, Linear Algebra Appl., 29 (1980), pp. 293–322.
[37] X. Pinel, *A Perturbed Two-level Preconditioner for the Solution of Three-dimensional Heterogeneous Helmholtz Problems with Applications to Geophysics*, Ph.D. thesis, Department of Mathematics, CERFACS, Toulouse, France, 2010.
[38] M. Robbé and M. Sadkane, *Exact and Inexact Breakdowns in Block Versions of FOM and GMRES Methods*, Technical Report, Département de Mathématiques, Université de Bretagne Occidentale, Brest, France, 2004; available online at http://www.math.univ-brest.fr/archives/recherche/prepub/Archives/2005/breakdowns.pdf.
[39] M. Robbé and M. Sadkane, *Exact and inexact breakdowns in the block GMRES method*, Linear Algebra Appl., 419 (2006), pp. 265–285.
[40] A. Ruhe, *Implementation aspects of band Lanczos algorithms for computation of eigenvalues of large sparse symmetric matrices*, Math. Comput., 33 (1979), pp. 680–687.
[41] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14 (1993), pp. 461–469.
[42] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.
[43] T. Sakurai, H. Tadano, and Y. Kuramashi, *Application of block Krylov subspace algorithms to the Wilson-Dirac equation with multiple right-hand sides in lattice QCD*, Comput. Phys. Commun., 181 (2010), pp. 113–117.
[44] P. Soudais, *Iterative solution methods of a 3-D scattering problem from arbitrary shaped multidielectric and multiconducting bodies*, IEEE Trans. Antennas and Propagation, 42 (1994), pp. 954–959.
[45] X. Sun and C. Bischof, *A basis-kernel representation of orthogonal matrices*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1184–1196.
[46] A. Toselli and O. Widlund, *Domain Decomposition Methods—Algorithms and Theory*, Springer Ser. Comput. Math. 34, Springer, New York, 2004.
[47] J. Virieux and S. Operto, *An overview of full waveform inversion in exploration geophysics*, Geophys., 74 (2009), pp. WCC127–WCC152.
[48] B. Vital, *Etude de quelques méthodes de résolution de problème linéaire de grande taille sur multiprocesseur*, Ph.D. thesis, Department of Computer Science, Université de Rennes, Rennes, France, 1990.
[49] R. Yu, E. de Sturler, and D. D. Johnson, *A Block Iterative Solver for Complex Non-Hermitian Systems Applied to Large-Scale Electronic-Structure Calculations*, Technical Report UIUCDCS-R-2002-2299, Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL, 2002.

Downloaded 11/04/13 to 194.199.19.9. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Bibliography

[1] J. A. Acebrón, A. Rodriguez-Rozas, and R. Spigler. "Domain decomposition solution of nonlinear two-dimensional parabolic problems by random trees". In: *J. Comp. Phys.* 228 (2009), pp. 5574–5591 (cit. on p. 111).

[2] J. A. Acebrón, A. Rodriguez-Rozas, and R. Spigler. "Efficient parallel solution of nonlinear parabolic partial differential equations by a probabilistic domain decomposition". In: *J. Sci. Comput.* 43.2 (2009), pp. 135–157 (cit. on p. 111).

[3] J. A. Acebrón, M. Busico, P. Lanucara, and R. Spigler. "Domain decomposition solution of elliptic boundary-value problems via Monte Carlo and Quasi-Monte Carlo methods". In: *SIAM J. Sci. Comput.* 27.2 (2005), pp. 135–157 (cit. on p. 111).

[4] M. Adams. "A low memory, highly concurrent multigrid algorithm". In: *ArXiv e-prints* (2012). arXiv: `1207.6720 [math.NA]` (cit. on p. 110).

[5] M. Adams, J. Brown, M. Knepley, and R. Samtaney. "Segmental refinement: a multigrid technique for data locality". In: *SIAM J. Sci. Comput.* (2015 (accepted)). arXiv:1406.7808 (cit. on p. 110).

[6] B. Aksoylu and H. Klie. "A family of physics-based preconditioners for solving elliptic equations on highly heterogeneous media". In: *Appl. Num. Math.* 59 (2009), pp. 1159–1186 (cit. on p. 66).

[7] P. R. Amestoy, I. S. Duff, and J. Y. L'Excellent. "Multifrontal parallel distributed symmetric and unsymmetric solvers". In: *Comput. Methods Appl. Mech. Engrg.* 184 (2000), pp. 501–520 (cit. on p. 9).

[8] P. R. Amestoy, I. S. Duff, J. Koster, and J. Y. L'Excellent. "A fully asynchronous multifrontal solver using distributed dynamic scheduling". In: *SIAM J. Matrix Anal. Appl.* 23 (1) (2001), pp. 15–41 (cit. on p. 9).

[9] P. R. Amestoy, A. Guermouche, J. Y. L'Excellent, and S. Pralet. "Hybrid scheduling for the parallel solution of linear systems". In: *Parallel Comput.* 32(2) (2006), pp. 136–156 (cit. on p. 9).

[10] P. R. Amestoy, C. Ashcraft, O. Boiteau, A. Buttari, J.-Y. L'Excellent, and C. Weisbecker. "Improving multifrontal methods by means of Block Low-Rank representations". In: *SIAM J. Sci. Comput.* 37.3 (2015), A1451–A1474 (cit. on p. 9).

[11] F. Aminzadeh, J. Brac, and T. Kunz. *3D Salt and Overthrust Models*. SEG/EAGE modeling series I. Society of Exploration Geophysicists, 1997 (cit. on p. 25).

[12] B. Andersson, U. Falk, I. Babuška, and T. von Petersdorff. "Reliable stress and fracture mechanics analysis of complex aircraft components using a *hp*–version FEM". In: *Int. J. Numer. Meth. Eng.* 38.13 (1995), pp. 2135–2163 (cit. on pp. 36, 41).

[13] M. Arioli. "A stopping criterion for the conjugate gradient algorithm in a finite element framework". In: *Numer. Math.* 97 (2004), pp. 1–24 (cit. on p. 110).

[14] M. Arioli, D. Loghin, and A. Wathen. "Stopping criteria for iterations in finite element methods". In: *Numer. Math.* 99 (2005), pp. 381–410 (cit. on p. 110).

[15] O. Axelsson. *Iterative Solution Methods*. Cambridge University Press, Cambridge, 1994 (cit. on pp. 1, 3, 71).

[16] O. Axelsson and P. S. Vassilevski. "Algebraic multilevel preconditioning methods. I". In: *Numer. Math.* 56 (1989), pp. 157–177 (cit. on p. 91).

[17] I. Babuška and B. Guo. "Approximation properties of the *hp*–version of the finite element method". In: *Comput. Methods Appl. Mech. Engrg.* 133 (1996), pp. 319–346 (cit. on pp. 33, 36, 41).

[18] S. Badia, A. F. Martin, and J. Principe. "A highly scalable parallel implementation of balancing domain decomposition by constraints". In: *SIAM J. Sci. Comput.* (2014), pp. C190–C218 (cit. on p. 96).

[19] S. Badia, A. F. Martin, and J. Principe. "Multilevel balancing domain decomposition at extreme scales". In: *SIAM J. Sci. Comput.* 38-1 (2016), pp. C22–C52 (cit. on p. 96).

[20] S. Badia, A. F. Martin, and J. Principe. "On the scalability of inexact balancing domain decomposition by constraints with overlapped coarse/fine corrections". In: *Parallel Comput.* 50 (2015), pp. 1–24 (cit. on p. 96).

[21] J. Baglama. "Augmented block Householder Arnoldi method". In: *Linear Algebra Appl.* 429 (2008), pp. 2315–2334 (cit. on p. 78).

[22] A. H. Baker, J. M. Dennis, and E. R. Jessup. "An efficient block variant of GMRES". In: *SIAM J. Sci. Comput.* 27 (2006), pp. 1608–1626 (cit. on p. 76).

[23] A. H. Baker, E. R. Jessup, and T. Manteuffel. "A technique for accelerating the convergence of restarted GMRES". In: *SIAM J. Matrix Anal. Appl.* 26.4 (2005), pp. 962–984 (cit. on p. 71).

[24] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition*. Philadelphia, PA: SIAM, 1994 (cit. on p. 1).

[25] A. Bayliss, C. I. Goldstein, and E. Turkel. "An iterative method for the Helmholtz equation". In: *J. Comp. Phys.* 49 (1983), pp. 443–457 (cit. on pp. 10, 13).

[26] A. Bellen and M. Zennaro. "Parallel algorithms for initial-value problems for difference and differential equations". In: *J. Comput. Appl. Math.* 25.3 (1989), pp. 341–350 (cit. on p. 107).

[27] M. Benzi. "Preconditioning techniques for large linear systems: a survey". In: *J. Comp. Phys.* 182 (2002), pp. 418–477 (cit. on p. 2).

[28] M. Benzi, G. H. Golub, and J. Liesen. "Numerical solution of saddle point problems". In: *Acta Numer.* 14.1 (2005), pp. 1–137 (cit. on p. 100).

[29] M. Benzi, A. Frommer, R. Nabben, and D. Szyld. "Algebraic theory of multiplicative Schwarz methods". In: *Numer. Math.* 89 (2001), pp. 605–639 (cit. on p. 3).

[30] J.-P. Berenger. "A perfectly matched layer for absorption of electromagnetic waves". In: *J. Comp. Phys.* 114 (1994), pp. 185–200 (cit. on p. 11).

[31] J.-P. Berenger. "Three-dimensional perfectly matched layer for absorption of electromagnetic waves". In: *J. Comp. Phys.* 127 (1996), pp. 363–379 (cit. on pp. 11, 13).

[32] I. Bermejo-Moreno, J. Bodart, and J. Larsson. "Scaling compressible flow solvers on the IBM Blue Gene/Q platform on up to 1.97 million cores". In: *Center for Turbulence Research, Annual Research Briefs* n/a (2013), pp. 343–358 (cit. on p. 106).

[33] I. Bermejo-Moreno, J. Bodart, J. Larsson, B. Barney, J. W. Nichols, and S. Jones. "Solving the Compressible Navier-Stokes Equations on Up to 1.97 Million Cores and 4.1 Trillion Grid Points". In: *Proceedings of the International Conference*

on High Performance Computing, Networking, Storage and Analysis. SC '13. Denver, Colorado: ACM, 2013, 62:1–62:10 (cit. on p. 106).

[34] F. Bernal and J. A. Acebrón. "A multigrid-like algorithm for Probabilistic Domain Decomposition". In: *ArXiv e-prints* (). arXiv: `1512.02818 [math.NA]` (cit. on p. 111).

[35] C. Bernardi and Y. Maday. "Spectral methods". In: *Handbook of Numerical Analysis, Vol. V, Part 2*. Amsterdam: North-Holland, 1997, pp. 209–485 (cit. on p. 36).

[36] R. Blaheta. "A multilevel method with correction by aggregation for solving discrete elliptic problems". In: *Appl. Math.* 31.5 (1986), pp. 365–378 (cit. on pp. 90, 91).

[37] M. Bollhöfer, M. J. Grote, and O. Schenk. "Algebraic multilevel preconditioner for the solution of the Helmholtz equation in heterogeneous media". In: *SIAM J. Sci. Comput.* 31 (2009), pp. 3781–3805 (cit. on pp. 10, 15).

[38] D. Braess. "Towards algebraic multigrid for elliptic problems of second order". In: *Computing* 55.4 (1995), pp. 379–393 (cit. on pp. 90, 91).

[39] A. Brandt. "A multi-level adaptative solution to boundary-value problems". In: *Math. Comp.* 31 (1977), pp. 333–390 (cit. on pp. 3, 8, 105, 110).

[40] A. Brandt and O. E. Livne. *Multigrid Techniques: 1984 Guide with Applications to Fluid Dynamics*. Revised Edition. SIAM, Philadelphia, 2011 (cit. on pp. 3, 110).

[41] A. Brandt and I. Livshits. "Wave-ray multigrid method for standing wave equations". In: *Electron. Trans. Numer. Anal.* 6 (1997), pp. 162–181 (cit. on p. 10).

[42] A. Brandt, S. F. McCormick, and J. W. Ruge. *Algebraic multigrid (AMG) for automatic multigrid solutions with application to geodetic computations*. Technical Report. Institute for Computational Studies, Fort Collins, CO, 1982 (cit. on pp. 89, 90).

[43] A. Brandt and S. Ta'asan. "Multigrid method for nearly singular and slightly indefinite problems". In: *Multigrid Methods II*. Ed. by W. Hackbusch and U. Trottenberg. Springer-Verlag, 1986, pp. 99–121 (cit. on p. 10).

[44] M. Brezina, R. D. Falgout, S. P. McLachlan, T.A. Manteuffel, S. F. McCormick, and J. W. Ruge. "Adaptive smoothed aggregation (αSA)". In: *SIAM J. Sci. Comput.* 25.6 (2004), pp. 1896–1920 (cit. on p. 91).

[45] W. L. Briggs, V. E. Henson, and S. F. McCormick. *A Multigrid Tutorial*. SIAM, 2000 (cit. on p. 3).

[46] P. N. Brown and H. F. Walker. "GMRES on (nearly) singular systems". In: *SIAM J. Matrix Anal. Appl.* 18 (1997), pp. 37–51 (cit. on p. 69).

[47] C. G. Broyden. "A class of methods for solving nonlinear simultaneous equations". In: *Math. Comp.* 19 (1965), pp. 577–593 (cit. on p. 103).

[48] C. G. Broyden. "A new method of solving nonlinear simultaneous equations". In: *The Computer Journal* 12.1 (1969), pp. 94–99 (cit. on p. 103).

[49] B. V. Budaev and D. B. Bogy. "Novel solutions of the Helmholtz equation and their application to diffraction". In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 463.2080 (2007), pp. 1005–1027 (cit. on p. 111).

[50] B. V. Budaev and D. B. Bogy. "Probabilistic solutions of the Helmholtz equation". In: *The Journal of the Acoustical Society of America* 109.5 (2001), pp. 2260–2262 (cit. on p. 111).

[51] V. Bulgakov. "Multi-level iterative technique and aggregation concept with semi-analytical preconditioning for solving boundary-value problems". In: *Comm. Numer. Methods Engrng.* 9 (1993), pp. 649–657 (cit. on p. 90).

[52] P. A. Businger and G. Golub. "Linear least squares solutions by Householder transformations". In: *Numer. Math.* 7 (1965), pp. 269–276 (cit. on p. 81).

[53] E. J. Bylaska, J. Q. Weare, and J. H. Weare. "Extending molecular simulation time scales: Parallel in time integrations for high-level quantum chemistry and complex force representations". In: *The Journal of Chemical Physics* 139.7 (2013), p. 074114 (cit. on p. 106).

[54] R. Byrd, J. Nocedal, and R. Schnabel. "Representations of quasi-Newton matrices and their use in limited memory methods". In: *Mathematical Programming* 63 (1994), pp. 129–156 (cit. on p. 105).

[55] X.-C. Cai and D. E. Keyes. "Nonlinearly preconditioned inexact Newton algorithms". In: *SIAM J. Sci. Comput.* 24 (2002), pp. 183–200 (cit. on p. 105).

[56] X.-C. Cai and M. Sarkis. "A restricted additive Schwarz preconditioner for general sparse linear systems". In: *SIAM J. Sci. Comput.* 21 (1999), pp. 792–797 (cit. on p. 97).

[57] H. Calandra, S. Gratton, R. Lago, X. Vasseur, and L. M. Carvalho. "A modified block flexible GMRES method with deflation at each iteration for the solution of non-Hermitian linear systems with multiple right-hand sides". In: *SIAM J. Sci. Comput.* 35.5 (2013), S345–S367 (cit. on pp. 59, 84–86).

[58] H. Calandra, S. Gratton, X. Pinel, and X. Vasseur. "An improved two-grid preconditioner for the solution of three-dimensional Helmholtz problems in heterogeneous media". In: *Numer. Linear Algebra Appl.* 20 (2013), pp. 663–688 (cit. on pp. 7, 16, 20, 23–28, 32).

[59] H. Calandra, S. Gratton, J. Langou, X. Pinel, and X. Vasseur. "Flexible variants of block restarted GMRES methods with application to geophysics". In: *SIAM J. Sci. Comput.* 34.2 (2012), A714–A736 (cit. on pp. 25, 59, 78, 80, 82, 84–86).

[60] H. Calandra, S. Gratton, R. Lago, X. Pinel, and X. Vasseur. "Two-level preconditioned Krylov subspace methods for the solution of three-dimensional heterogeneous Helmholtz problems in seismics". In: *Numerical Analysis and Applications* 5.2 (2012), pp. 175–181 (cit. on p. 86).

[61] L. M. Carvalho, S. Gratton, R. Lago, and X. Vasseur. "A flexible Generalized Conjugate Residual method with inner orthogonalization and deflated restarting". In: *SIAM J. Matrix Anal. Appl.* 32.4 (2011), pp. 1212–1235 (cit. on pp. 59, 73, 74, 86, 97).

[62] P. Castillo, R. Rieben, and D. White. "FEMSTER: An object-oriented class library of high-order discrete differential forms". In: *ACM Trans. Math. Softw.* 31.4 (2005), pp. 425–457 (cit. on p. 55).

[63] A. Chapman and Y. Saad. "Deflated and augmented Krylov subspace techniques". In: *Numer. Linear Algebra Appl.* 4.1 (1997), pp. 43–66 (cit. on pp. 62, 63).

[64] P. Chartier and B. Philippe. "A parallel shooting technique for solving dissipative ODE's". In: *Computing* 51.3-4 (1993), pp. 209–236 (cit. on p. 108).

[65] Z. Chen, D. Cheng, and T. Wu. "A dispersion minimizing finite difference scheme and preconditioned solver for the 3D Helmholtz equation". In: *J. Comp. Phys.* 231 (2012), pp. 8152–8175 (cit. on pp. 12, 13, 15, 30).

[66] G. Cohen. *Higher-order Numerical Methods for Transient Wave Equations*. Springer Verlag, Heidelberg, 2002 (cit. on p. 12).

[67] L. Conen, V. Dolean, R. Krause, and F. Nataf. "A coarse space for heterogeneous Helmholtz problems based on the Dirichlet-to-Neumann operator". In: *J. Comput. Appl. Math.* 271 (2014), pp. 83–99 (cit. on p. 97).

[68] L. Conen, V. Dolean, R. Krause, and F. Nataf. "Addendum to A coarse space for heterogeneous Helmholtz problems based on the Dirichlet-to-Neumann operator". In: *J. Comput. Appl. Math.* 290 (2015), pp. 670–674 (cit. on p. 97).

[69] S. F. Mc Cormick and D. Quinlan. "Asynchronous multilevel adaptive methods for solving partial differential equations on multiprocessors: Performance results". In: *Parallel Comput.* 12 (1989), pp. 145–156 (cit. on p. 110).

[70] R. Croce, D. Ruprecht, and R. Krause. "Parallel-in-Space-and-Time simulation of the three-dimensional, unsteady Navier-Stokes equations for incompressible flow". In: *Modeling, Simulation and Optimization of Complex Processes – HPSC 2012*. Ed. by Hans Georg Bock, Xuan Phu Hoang, Rolf Rannacher, and Johannes P. Schlöder. Springer International Publishing, 2014, pp. 13–23 (cit. on p. 108).

[71] T. Cui, J. Xu, and C.-S. Zhang. "An error-resilient redundant subspace correction method". In: *ArXiv e-prints* (2013). arXiv: `1309.0212 [math.NA]` (cit. on p. 110).

[72] D. Darnell, R. B. Morgan, and W. Wilcox. "Deflation of eigenvalues for iterative methods in lattice QCD". In: *Nuclear Physics B - Proceedings Supplements* 129-130 (2004), pp. 856–858 (cit. on p. 66).

[73] T. A. Davis. *Direct Methods for Sparse Linear Systems*. SIAM, Philadelphia, 2006 (cit. on pp. 9, 76).

[74] F. Delbaen, J. Qiu, and S. Tang. "Forward-backward stochastic differential systems associated to Navier-Stokes equations in the whole space". In: *Stochastic Processes and their Applications* 125.7 (2015), pp. 2516 –2561 (cit. on p. 111).

[75] R. S. Dembo, S.C. Eisenstat, and T. Steihaug. "Inexact Newton methods". In: *SIAM J. Numer. Anal.* 19 (1982), pp. 400–402 (cit. on p. 105).

[76] G. B. Deng, J. Piquet, X. Vasseur, and M. Visonneau. "A new fully coupled method for computing turbulent flows". In: *Comput. Fluids* 30.4 (2001), pp. 445–472 (cit. on p. 8).

[77] Y. Diouane. "Globally Convergent Evolution Strategies with Application to an Earth Imaging Problem in Geophysics". PhD thesis. CERFACS, Toulouse, France, 2014 (cit. on p. 32).

[78] Y. Diouane, S. Gratton, X. Vasseur, L. N. Vicente, and H. Calandra. "A parallel evolution strategy for an Earth imaging problem in geophysics". In: *Optimization and Engineering* 17.1 (2016), pp. 3–26 (cit. on p. 32).

[79] C. Dohrmann. "A preconditioner for substructuring based on constrained energy minimization". In: *SIAM J. Sci. Comput.* 25 (2003), pp. 246–258 (cit. on p. 96).

[80] V. Dolean, P. Jolivet, and F. Nataf. *An Introduction to Domain Decomposition Methods: Algorithms, Theory and Parallel Implementation.* SIAM, 2015 (cit. on pp. 3, 55, 95).

[81] M. Dryja, M. V. Sarkis, and O. B. Widlund. "Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions". In: *Numer. Math.* 72.3 (1996), pp. 313–348 (cit. on pp. 38, 53).

[82] M. Dryja and O. B. Widlund. "Schwarz methods of Neumann-Neumann type for three-dimensional elliptic finite element problems". In: *Comm. Pure Appl. Math.* 48.2 (1995), pp. 121–155 (cit. on pp. 38, 53).

[83] I. S. Duff, A. M. Erisman, and J. K. Reid. *Direct Methods for Sparse Matrices.* Oxford University Press, 1989 (cit. on p. 76).

[84] I. S. Duff and J. K. Reid. "The multifrontal solution of unsymmetric sets of linear systems". In: *SIAM J. Sci. Comput.* 5 (1984), pp. 633–641 (cit. on p. 9).

[85] W. Edwards, L. Tuckerman, R. Friesner, and D. Sorensen. "Krylov methods for the incompressible Navier-Stokes equations". In: *J. Comp. Phys.* 110 (1994), pp. 82–102 (cit. on p. 109).

[86] M. Eiermann and O. G. Ernst. "Geometric aspects of the theory of Krylov subspace methods". In: *Acta Numer.* 10 (2001), pp. 251–312 (cit. on p. 59).

[87] M. Eiermann, O. G. Ernst, and O. Schneider. "Analysis of acceleration strategies for restarted minimal residual methods." In: *J. Comput. Appl. Math.* 123 (2000), pp. 261–292 (cit. on pp. 59, 71, 72).

[88] T. Eirola and O. Nevanlinna. "Accelerating with rank-one updates". In: *Linear Algebra Appl.* 121 (1989), pp. 511–520 (cit. on p. 103).

[89] S. C. Eisenstat, H. C. Elman, and M. H. Schultz. "Variational iterative methods for nonsymmetric systems of linear equations". In: *SIAM J. Numer. Anal.* 20.2 (1983), pp. 345–357 (cit. on pp. 71, 104).

[90] H. Elman, D. Silvester, and A. Wathen. *Finite Elements and Fast Iterative Solvers: with applications in incompressible fluid dynamics*. Second edition. Oxford University Press, 2014 (cit. on pp. 1, 95).

[91] H. Elman, O. Ernst, D. O'Leary, and M. Stewart. "Efficient iterative algorithms for the stochastic finite element method with application to acoustic scattering". In: *Comput. Methods Appl. Mech. Engrg.* 194.1 (2005), pp. 1037–1055 (cit. on p. 77).

[92] H. C. Elman, O. G. Ernst, and D. P. O'Leary. "A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations". In: *SIAM J. Sci. Comput.* 23 (2001), pp. 1291–1315 (cit. on p. 16).

[93] B. Engquist and L. Ying. "Sweeping preconditioner for the Helmholtz equation: hierarchical matrix representation". In: *Comm. Pure Appl. Math.* 64 (2011), pp. 697–735 (cit. on p. 15).

[94] B. Engquist and L. Ying. "Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers". In: *Multiscale Modeling and Simulation* 9 (2011), pp. 686–710 (cit. on p. 15).

[95] Y. A. Erlangga. "A Robust and Efficient Iterative Method for the Numerical Solution of the Helmholtz Equation". PhD thesis. Delft University of Technology, The Netherlands, 2005 (cit. on p. 10).

[96] Y. A. Erlangga. "Advances in iterative methods and preconditioners for the Helmholtz equation". In: *Archives of Computational Methods in Engineering* 15 (2008), pp. 37–66 (cit. on pp. 7, 9, 15).

[97] Y. A. Erlangga and R. Nabben. "Deflation and balancing preconditioners for Krylov subspace methods applied to nonsymmetric matrices". In: *SIAM J. Matrix Anal. Appl.* 30.2 (2008), pp. 684–699 (cit. on pp. 10, 70, 71).

[98] Y. A. Erlangga, C. W. Oosterlee, and C. Vuik. "A novel multigrid based preconditioner for heterogeneous Helmholtz problems". In: *SIAM J. Sci. Comput.* 27 (2006), pp. 1471–1492 (cit. on pp. 10, 13, 15, 22).

[99] Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. "On a class of preconditioners for solving the Helmholtz equation". In: *Appl. Num. Math.* 50 (2004), pp. 409–425 (cit. on pp. 10, 13, 14).

[100] O. G. Ernst and M. J. Gander. "Why it is difficult to solve Helmholtz problems with classical iterative methods". In: *Numerical Analysis of Multiscale Problems*. Ed. by I. Graham, T. Hou, O. Lakkis, and R. Scheichl. Vol. 83. Lecture Notes

in Computational Science and Engineering. Berlin Heidelberg: Springer-Verlag, 2012, pp. 325–361 (cit. on pp. 4, 7, 9, 15).

[101] C. Farhat, A. Macedo, and M. Lesoinne. "A two-level domain decomposition method for the iterative solution of high frequency exterior Helmholtz problems". In: *Numer. Math.* 85 (2000), pp. 283–308 (cit. on p. 9).

[102] C. Farhat, K. Pierson, and M. Lesoinne. "The second generation FETI methods and their application to the parallel solution of large-scale linear and geometrically non-linear structural analysis problems". In: *Comput. Methods Appl. Mech. Engrg.* 184 (2000), pp. 333–374 (cit. on p. 96).

[103] C. Farhat and F.-X. Roux. "A method of finite element tearing and interconnecting and its parallel solution algorithm". In: *Int J. Numerical Methods in Engineering* 32 (1991), pp. 1205–1227 (cit. on pp. 9, 34, 43).

[104] C. Farhat and F.-X. Roux. "Implicit parallel processing in structural mechanics". In: *Computational Mechanics Advances.* Ed. by J. Tinsley Oden. Vol. 2 (1). North-Holland, 1994, pp. 1–124 (cit. on p. 43).

[105] C. Farhat, M. Lesoinne, P. Le Tallec, K. Pierson, and D. Rixen. "FETI-DP: a dual-primal unified FETI method. I. A faster alternative to the two-level FETI method". In: *Internat. J. Numer. Methods Engrg.* 50.7 (2001), pp. 1523–1544 (cit. on p. 55).

[106] P. F. Fischer, F. Hecht, and Y. Maday. "A parareal in time semi-implicit approximation of the Navier-Stokes equations". In: *Domain Decomposition Methods in Science and Engineering.* Ed. by Ralf Kornhuber and et al. Vol. 40. Lecture Notes in Computational Science and Engineering. Berlin: Springer, 2005, pp. 433–440 (cit. on p. 108).

[107] D. Fokkema. "Subspace Methods for Linear, Nonlinear and Eigen Problems". PhD thesis. University of Utrecht, The Netherlands, 1996 (cit. on p. 72).

[108] R. W. Freund and M. Malhotra. "A block QMR algorithm for non-Hermitian linear systems with multiple right-hand sides". In: *Linear Algebra Appl.* 254 (1997), pp. 119–157 (cit. on p. 77).

[109] R. Friesner, L. Tuckerman, B. Dornblaser, and T. Russo. "A method for exponential propagation of large systems of stiff nonlinear differential equations". In: *J. Sci. Comput.* 4 (1989), pp. 327–354 (cit. on p. 109).

[110] A. Frommer, A. Nobile, and P. Zingler. *Deflation and Flexible SAP Preconditioning of GMRES in Lattice QCD Simulation.* Technical Report BUW-IMACM 12/11. Department of Mathematics: University of Wuppertal, 2012 (cit. on p. 66).

[111] A. Frommer and D. Szyld. "An algebraic convergence theory for restricted additive Schwarz methods using weighted max norms". In: *SIAM J. Numer. Anal.* 39.2 (2001), pp. 463–479 (cit. on p. 3).

[112] A. Frommer and D. Szyld. "On asynchronous iterations". In: *J. Comput. Appl. Math.* 123 (2000), pp. 201–216 (cit. on p. 110).

[113] A. Frommer and D. Szyld. "Weighted max norms, splittings, and overlapping additive Schwarz iterations". In: *Numer. Math.* 83 (1999), pp. 259–278 (cit. on p. 3).

[114] M. Gander, I. G. Graham, and E. A. Spence. "Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: What is the largest shift for which wavenumber-independent convergence is guaranteed?" In: *Numer. Math.* 131 (2015), pp. 567–614 (cit. on p. 15).

[115] M. J. Gander. "50 years of time parallel time integration". In: *Multiple Shooting and Time Domain Decomposition.* Springer Verlag, 2015, pp. 69–114 (cit. on pp. 106, 108).

[116] M. J. Gander and S. Güttel. "ParaExp: A parallel integrator for linear initial-value problems". In: *SIAM J. Sci. Comput.* 35.2 (2013), pp. C123–C142 (cit. on pp. 108, 109).

[117] M. J. Gander, F. Magoulès, and F. Nataf. "Optimized Schwarz methods without overlap for the Helmholtz equation". In: *SIAM J. Sci. Comput.* 24 (2002), pp. 38–60 (cit. on p. 10).

[118] I. Garrido, M. S. Espedal, and G. E. Fladmark. "A convergent algorithm for time parallelization applied to reservoir simulation". In: *Domain Decomposition Methods in Science and Engineering.* Ed. by Timothy J. Barth and al. Vol. 40. Lecture Notes in Computational Science and Engineering. Springer Berlin Heidelberg, 2005, pp. 469–476 (cit. on p. 108).

[119] A. Gaul. "Recycling Krylov Subspace Methods for Sequences of Linear Systems". PhD Thesis. Technische Universität Berlin, Germany, 2014 (cit. on pp. 87, 97, 103, 104).

[120] A. Gaul and N. Schlömer. "Preconditioned recycling Krylov subspace methods for self-adjoint problems". In: *Electron. Trans. Numer. Anal.* 44 (2015), pp. 522–547 (cit. on pp. 97, 103).

[121] A. Gaul, M. Gutknecht, J. Liesen, and R. Nabben. "A framework for deflated and augmented Krylov subspace methods". In: *SIAM J. Matrix Anal. Appl.* 34 (2013), pp. 495–518 (cit. on pp. 67, 104).

[122] A. Gaul, M. Gutknecht, J. Liesen, and R. Nabben. *Deflated and augmented Krylov subspace methods: Basic facts and a breakdown-free deflated MINRES*. Preprint Preprint 759. TU Berlin: DFG Research Center MATHEON, 2011 (cit. on pp. 67, 69).

[123] M. B. van Gijzen, Y. A. Erlangga, and C. Vuik. "Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian". In: *SIAM J. Sci. Comput.* 29 (2007), pp. 1942–1958 (cit. on pp. 10, 15).

[124] L. Giraud, S. Gratton, X. Pinel, and X. Vasseur. "Flexible GMRES with deflated restarting". In: *SIAM J. Sci. Comput.* 32.4 (2010), pp. 1858–1878 (cit. on pp. 59, 64–66, 86).

[125] B. Gmeiner, U. Rüde, H. Stengel, C. Waluga, and B. Wohlmuth. "Performance and scalability of hierarchical hybrid multigrid solvers for Stokes systems". In: *SIAM J. Sci. Comput.* 37.2 (2015), pp. C143–C168 (cit. on p. 110).

[126] B. Gmeiner, U. Rüde, H. Stengel, C. Waluga, and B. Wohlmuth. "Towards textbook efficiency for parallel multigrid". In: *Numerical Mathematics: Theory, Methods and Applications* 8.01 (2015), pp. 22–46 (cit. on p. 110).

[127] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Third edition. The Johns Hopkins University Press, 1996 (cit. on pp. 79, 81).

[128] S. Gratton, A. Sartenaer, and J. Tshimanga. "On a class of limited memory preconditioners for large scale linear systems with multiple right-hand sides". In: *SIAM J. Opt.* 21.3 (2011), pp. 912–935 (cit. on p. 98).

[129] S. Gratton, D. Titley-Peloquin, P. Toint, and J. Tshimanga Ilunga. "Differentiating the method of conjugate gradients". In: *SIAM J. Matrix Anal. Appl.* 35 (2014), pp. 110–126 (cit. on p. 104).

[130] S. Gratton, S. Mercier, N. Tardieu, and X. Vasseur. *Limited memory preconditioners for symmetric indefinite problems with application to structural mechanics*. Technical Report TR/PA/15/48. CERFACS, Toulouse, France, 2015 (cit. on pp. 98–102).

[131] S. Gratton, P. Hénon, P. Jiránek, and X. Vasseur. *Reducing complexity of algebraic multigrid by aggregation*. Technical Report TR/PA/14/18. CERFACS, Toulouse, France, 2014 (cit. on p. 95).

[132] S. Gratton, P. Hénon, P. Jiránek, and X. Vasseur. "Reducing complexity of algebraic multigrid by aggregation". In: *Numer. Linear Algebra Appl.* 23 (2016), pp. 501–518 (cit. on pp. 92–94).

[133] A. Greenbaum. *Iterative Methods for Solving Linear Systems*. SIAM, Philadelphia, 1997 (cit. on pp. 1, 3).

[134] W. Gropp, E. Lusk, and A. Skjellum. *Using MPI: Portable Parallel Programming with the Message-Passing Interface*. MIT Press, 1999 (cit. on p. 26).

[135] A. El Guennouni, K. Jbilou, and H. Sadok. "A block version of BICGSTAB for linear systems with multiple right-hand sides". In: *Electron. Trans. Numer. Anal.* 16 (2003), pp. 129–142 (cit. on p. 77).

[136] M. H. Gutknecht. "Block Krylov space methods for linear systems with multiple right-hand sides: an introduction". In: *Modern Mathematical Models, Methods and Algorithms for Real World Systems*. Ed. by A.H. Siddiqi, I.S. Duff, and O. Christensen. New Delhi, India: Anamaya Publishers, 2006, pp. 420–447 (cit. on pp. 58, 76, 77, 79, 81).

[137] M. H. Gutknecht. "Deflated and augmented Krylov subspace methods: A framework for deflated BiCG and related solvers". In: *SIAM J. Matrix Anal. Appl.* 35 (2014), pp. 1444–1466 (cit. on p. 104).

[138] M. H. Gutknecht. "Spectral deflation in Krylov solvers: A theory of coordinate space based methods". In: *Electron. Trans. Numer. Anal.* 39 (2012), pp. 156–185 (cit. on pp. 67–70).

[139] M. H. Gutknecht and T. Schmelzer. "The block grade of a block Krylov space". In: *Linear Algebra Appl.* 430.1 (2009), pp. 174–185 (cit. on p. 76).

[140] W. Hackbusch. *Iterative Solution of Large Sparse Systems of Equations*. Vol. 95. Applied mathematical sciences. New York, NY: Springer, 1994 (cit. on p. 1).

[141] W. Hackbusch. *Multi-Grid methods and Applications*. Second edition. Springer, 2003 (cit. on p. 3).

[142] G. Hager, J. Treibig, J. Habich, and G. Wellein. "Exploring performance and power properties of modern multicore chips via simple machine models". In: *Concurrency Computat.: Pract. Exper.* 28 (2016), pp. 189–210 (cit. on p. 110).

[143] I. Harari and E. Turkel. "Accurate finite difference methods for time-harmonic wave propagation". In: *J. Comp. Phys.* 119 (1995), pp. 252–270 (cit. on pp. 13, 29).

[144] L. Hart and S. F. Mc Cormick. "Asynchronous multilevel adaptive methods for solving partial differential equations on multiprocessors: Basic ideas". In: *Parallel Comput.* 12 (1989), pp. 131–144 (cit. on p. 110).

[145] V. E. Henson and U. M Yang. "*BoomerAMG*: A parallel algebraic multigrid solver and preconditioner". In: *Appl. Numer. Math.* 41 (2002), pp. 155–177 (cit. on p. 92).

[146] J. Hicken and D. Zingg. "A simplified and flexible variant of GCROT for solving nonsymmetric linear systems". In: *SIAM J. Sci. Comput.* 32.3 (2010), pp. 1672–1694 (cit. on p. 74).

[147] N. J. Higham. *Functions of Matrices: Theory and Computation*. Philadelphia, PA, USA: SIAM, 2008 (cit. on pp. 79, 98, 109).

[148] Y. P. Hong and C. T. Pan. "Rank revealing QR factorizations and the singular value decomposition". In: *Math. Comp.* 58 (1992), pp. 213–232 (cit. on p. 81).

[149] M. Huber, B. Gmeiner, U. Rüde, and B. Wohlmuth. "Resilience for multigrid software at the extreme scale". In: *ArXiv e-prints* (2015). arXiv: `1506.06185 [cs.MS]` (cit. on p. 110).

[150] P. Jolivet. "Méthodes de décomposition de domaine. Application au calcul haute performance." PhD thesis. Université de Grenoble, France, 2014 (cit. on p. 97).

[151] D. Kalchev, C. Ketelsen, and P. S. Vassilevski. "Two-level adaptive algebraic multigrid for a sequence of problems with slowly varying random coefficients". In: *SIAM J. Sci. Comput.* 35.6 (2015), B1215–B1234 (cit. on p. 95).

[152] G. E. Karniadakis and S. Sherwin. *Spectral/hp Element Methods for CFD*. Oxford University Press, 1999 (cit. on pp. 33, 34).

[153] R. Kirby. "From functional analysis to iterative methods". In: *SIAM Rev.* 52.2 (2010), pp. 269–293 (cit. on p. 1).

[154] A. Klawonn. "Block-triangular preconditioners for saddle point problems with a penalty term". In: *SIAM J. Sci. Comput.* 19.1 (1998), pp. 172–184 (cit. on p. 95).

[155] A. Klawonn, L. Pavarino, and O. Rheinbach. "Spectral element FETI-DP and BDDC preconditioners with multi-element subdomains". In: *Comput. Methods Appl. Mech. Engrg.* 198 (2008), pp. 511–523 (cit. on p. 55).

[156] A. Klawonn and O. Rheinbach. "Deflation, projector preconditioning, and balancing in iterative substructuring methods: connections and new results". In: *SIAM J. Sci. Comput.* 34.1 (2012), A459–A484 (cit. on p. 34).

[157] A. Klawonn and O. Rheinbach. "Highly scalable parallel domain decomposition methods with an application to biomechanics". en. In: *ZAMM Zeitschrift für Angewandte Mathematik und Mechanik* 90.1 (2010), pp. 5–32 (cit. on p. 96).

[158] A. Klawonn and O. B. Widlund. "FETI and Neumann-Neumann iterative substructuring methods: connections and new results". In: *Comm. Pure Appl. Math.* 54.1 (2001), pp. 57–90 (cit. on pp. 43–47).

[159] A. Klawonn, O. B. Widlund, and M. Dryja. "Dual-Primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients". In: *SIAM J. Numer. Anal.* 40.1 (2002), pp. 159–179 (cit. on p. 55).

[160] H. Klie, M.F. Wheeler, T. Clees, and K. Stüben. *Deflation AMG Solvers for Highly Ill-Conditioned Reservoir Simulation Problems*. Paper SPE 105820 presented at the 2007 SPE Reservoir Simulation Symposium, Houston, TX, Feb. 28-30 2007. 2007 (cit. on p. 66).

[161] G. L. Kooij, M. A. Botchev, and B. J. Geurts. "A block Krylov subspace implementation of the time-parallel Paraexp method and its extension for nonlinear partial differential equations". arXiv:1509.04567 [math.NA]. 2015 (cit. on p. 109).

[162] R. Lago. "A Study on Block Flexible Iterative Solvers with Application to Earth Imaging Problem in Geophysics". PhD thesis. CERFACS, Toulouse, France, 2013 (cit. on p. 32).

[163] A. L. Laird and M. B. Giles. *Preconditioned iterative solution of the 2D Helmholtz equation*. Technical Report Report NA-02/12. Oxford University Computing Laboratory, 2002 (cit. on p. 10).

[164] J. Langou. "Iterative Methods for Solving Linear Systems with Multiple Right-Hand Sides". PhD thesis. CERFACS, Toulouse, France, 2003 (cit. on pp. 76, 81).

[165] A. Lecerf. *Approche multisimulation pour l'amélioration des performances d'un solveur, analyse d'un algorithme autorisant la parallélisation en temps*. Master thesis report, ENSEEIHT, Toulouse. 2014 (cit. on p. 109).

[166] J. Liesen and Z. Strakoš. *Krylov Subspace Methods - Principles and Analysis*. Oxford University Press, 2013 (cit. on p. 3).

[167] J.-L. Lions, Y. Maday, and G. Turinici. "A parareal in time discretization of PDE's". In: *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* 332 (2001), pp. 661–668 (cit. on p. 106).

[168] F. Liu and L. Ying. "Additive sweeping preconditioner for the Helmholtz equation". In: *ArXiv e-prints* (2015). arXiv: `1504.04058 [math.NA]` (cit. on pp. 11, 31).

[169] F. Liu and L. Ying. "Recursive sweeping preconditioner for the 3D Helmholtz equation". In: *ArXiv e-prints* (2015). arXiv: `1502.07266 [math.NA]` (cit. on pp. 11, 31).

[170] M. Magolu Monga Made, R. Beauwens, and G. Warzee. "Preconditioning of discrete Helmholtz operators perturbed by a diagonal complex matrix". In: *Commun. Numer. Method Eng.* 11 (2000), pp. 801–817 (cit. on p. 10).

[171] J. Málek and Z. Strakoš. *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs*. SIAM, 2015 (cit. on pp. 1, 110).

[172] J. Mandel and M. Brezina. "Balancing Domain Decomposition for problems with large jumps in coefficients". In: *Math. Comp.* 65 (1996), pp. 1387–1401 (cit. on pp. 34, 37, 38, 53).

[173] J. Mandel and B. Sousedik. "BDDC and FETI-DP under minimalist assumptions". In: *Computing* 81 (2007), pp. 269–280 (cit. on p. 96).

[174] J. Mandel, B. Sousedik, and C. Dohrmann. "Multispace and multilevel BDDC". In: *Computing* 83.2 (2008), pp. 55–85 (cit. on p. 96).

[175] J. Mandel and R. Tezaur. "Convergence of a substructuring method with Lagrange multipliers". In: *Numer. Math.* 73 (1996), pp. 473–487 (cit. on p. 43).

[176] K. Mardal and R. Winther. "Preconditioning discretizations of systems of partial differential equations". In: *Numer. Linear Algebra Appl.* 18 (2011), pp. 1–40 (cit. on p. 1).

[177] S. F. McCormick. *Multigrid Methods*. SIAM, Philadelphia, 1987 (cit. on p. 3).

[178] J. M. Melenk and C. Schwab. "*hp*–FEM for reaction–diffusion equations. I: Robust exponential convergence". In: *SIAM J. Numer. Anal.* 35 (1998), pp. 1520–1557 (cit. on p. 36).

[179] S. Mercier. "Fast Nonlinear Solvers in Structural Mechanics". PhD thesis. University Paul Sabatier, Toulouse, France, 2015 (cit. on pp. 100, 103).

[180] G. Meurant. *Computer Solution of Large Linear Systems*. North-Holland, 1999 (cit. on p. 1).

[181] A. Moiola and E. Spence. "Is the Helmholtz equation really sign-indefinite ?" In: *SIAM Rev.* 56 (2014), pp. 274–312 (cit. on p. 9).

[182] J. L. Morales and J. Nocedal. "Automatic preconditioning by limited memory Quasi-Newton updating". In: *SIAM J. Opt.* 10.4 (2000), pp. 1079–1096 (cit. on pp. 98, 105, 108).

[183] R. B. Morgan. "A restarted GMRES method augmented with eigenvectors". In: *SIAM J. Matrix Anal. Appl.* 16 (1995), pp. 1154–1171 (cit. on pp. 63, 66).

[184] R. B. Morgan. "GMRES with deflated restarting". In: *SIAM J. Sci. Comput.* 24.1 (2002), pp. 20–37 (cit. on pp. 63, 66).

[185] R. B. Morgan. "Implicitly restarted GMRES and Arnoldi methods for non-symmetric systems of equations". In: *SIAM J. Matrix Anal. Appl.* 21.4 (2000), pp. 1112–1135 (cit. on p. 63).

[186] O. Mula. "Some Contributions Towards the Parallel Simulation of Time Dependent Neutron Transport and the Integration of Observed Data in Real Time". PhD thesis. University Pierre et Marie Curie, Paris VI, France, 2014 (cit. on pp. 106, 108).

[187] A. Muresan and Y. Notay. "Analysis of aggregation-based multigrid". In: *SIAM J. Sci. Comput.* 30.2 (2008), pp. 1082–1103 (cit. on p. 91).

[188] F. Nataf, H. Xiang, V. Dolean, and N. Spillane. "A coarse space construction based on local Dirichlet-to-Neumann maps". In: *SIAM J. Sci. Comput.* 33.4 (2011), pp. 1623–1642 (cit. on p. 97).

[189] J. Nečas and I. Hlavácek. *Mathematical Theory of Elastic and Elastoplastic Bodies*. Elsevier, 1980 (cit. on p. 100).

[190] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Ser. Oper. Res., Springer-Verlag, Berlin, 1999 (cit. on pp. 98, 105).

[191] Y. Notay. "An aggregation-based multigrid method". In: *Electron. Trans. Numer. Anal.* 37 (2010), pp. 123–146 (cit. on pp. 90, 91).

[192] Y. Notay and A. Napov. "A massively parallel solver for discrete Poisson-like problems". In: *J. Comp. Phys.* 281 (2015), pp. 237–250 (cit. on p. 90).

[193] Y. Notay and P. S. Vassilevski. "Recursive Krylov-based multigrid cycles". In: *Numer. Linear Algebra Appl.* 15 (2008), pp. 473–487 (cit. on pp. 18, 91).

[194] M. Olshankii and E. Tyrtyshnikov. *Iterative Methods for Linear Systems - Theory and Applications*. Philadelphia, PA, USA: SIAM, 2014 (cit. on pp. 1, 3).

[195] S. Operto, J. Virieux, P. R. Amestoy, J.-Y. L'Excellent, L. Giraud, and H. Ben Hadj Ali. "3D finite-difference frequency-domain modeling of visco-acoustic wave propagation using a massively parallel direct solver: A feasibility study". In: *Geophysics* 72-5 (2007), pp. 195–211 (cit. on p. 12).

[196] C. C. Paige, B. N. Parlett, and H. A. van der Vorst. "Approximate solutions and eigenvalue bounds from Krylov subspaces". In: *Numer. Linear Algebra Appl.* 2 (1995), pp. 115–134 (cit. on p. 63).

[197] M. Parks, E. de Sturler, G. Mc Key, D. D. Johnson, and S. Maiti. "Recycling Krylov subspaces for sequences of linear systems". In: *SIAM J. Sci. Comput.* 28.5 (2006), pp. 1651–1674 (cit. on pp. 72, 87, 97).

[198] L. F. Pavarino. "Neumann-Neumann algorithms for spectral elements in three dimensions". In: *RAIRO Mathematical Modelling and Numerical Analysis* 31 (1997), pp. 471–493 (cit. on pp. 34, 38, 39, 53, 54).

[199] J. Pestana and A. Wathen. "Natural preconditioning and iterative methods for saddle point systems". In: *SIAM Rev.* 57 (2015), pp. 71–91 (cit. on p. 1).

[200] X. Pinel. "A Perturbed Two-level Preconditioner for the Solution of Three-dimensional Heterogeneous Helmholtz Problems with Applications to Geophysics". PhD thesis. CERFACS, Toulouse, France, 2010 (cit. on pp. 12–14, 16, 17, 24, 25, 32).

[201] J. Piquet and X. Vasseur. "A non-standard multigrid method with flexible multiple semi-coarsening for the numerical solution of the pressure equation in a Navier-Stokes solver". In: *Numer. Algorithms* 4 (2000), pp. 333–355 (cit. on p. 8).

[202] J. Piquet and X. Vasseur. "Multigrid preconditioned Krylov subspace methods for three-dimensional numerical solutions of the incompressible Navier-Stokes equations". In: *Numer. Algorithms* 1-2 (1998), pp. 1–32 (cit. on p. 8).

[203] J. Poulson, B. Engquist, S. Li, and L. Ying. "A parallel sweeping preconditioner for heterogeneous 3D Helmholtz equations". In: *SIAM J. Sci. Comput.* 35 (2013), pp. C194–C212 (cit. on p. 9).

[204] A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations*. Oxford Science Publications, 1999 (cit. on p. 3).

[205] T. Rees and C. Greif. "A preconditioner for linear systems arising from interior point optimization methods". In: *SIAM J. Sci. Comput.* 29.5 (2007), pp. 1992–2007 (cit. on p. 101).

[206] O. Rheinbach. "Parallel iterative substructuring in structural mechanics". en. In: *Archives of Computational Methods in Engineering* 16.4 (2009), pp. 425–463 (cit. on p. 96).

[207] C. D. Riyanti, A. Kononov, Y. A. Erlangga, R.-E. Plessix, W. A. Mulder, C. Vuik, and C. W. Oosterlee. "A parallel multigrid-based preconditioner for the 3D heterogeneous high-frequency Helmholtz equation". In: *J. Comp. Phys.* 224 (2007), pp. 431–448 (cit. on pp. 10, 15).

[208] C. Rodrigo, F. Gaspar, C. W. Oosterlee, and I. Yavneh. "Accuracy measures and Fourier analysis for the full multigrid algorithm". In: *SIAM J. Sci. Comput.* 32 (2010), pp. 3108–3129 (cit. on p. 3).

[209] A. Rodriguez-Rozas. "Highly Efficient Probabilistic-Based Numerical Algorithms for Solving Partial Differential Equations on Massively Parallel Computers". PhD thesis. Instituto Superior Tecnico (Lisbon), Portugal, 2012 (cit. on p. 111).

[210] S. Röllin and W. Fichtner. "Improving the accuracy of GMRes with deflated restarting". In: *SIAM J. Sci. Comput.* 30.1 (2007), pp. 232–245 (cit. on p. 66).

[211] U. Rüde. "New mathematics for extreme-scale computational science ?" In: *SIAM News* 48.5 (2015), pp. 7–8 (cit. on p. 110).

[212] J. W. Ruge and K. Stüben. "Algebraic Multigrid". In: *Multigrid Methods*. Ed. by S. F. McCormick. SIAM: Philadelphia, PA, 1987, pp. 73–130 (cit. on pp. 89, 90).

[213] A. Ruhe. "Implementation aspects of band Lanczos algorithms for computation of eigenvalues of large sparse symmetric matrices". In: *Math. Comp.* 33.146 (1979), pp. 680–687 (cit. on p. 78).

[214] M. J. Ruijter and C. W. Oosterlee. "A Fourier cosine method for an efficient computation of solutions to BSDE". In: *SIAM J. Sci. Comput.* 37.2 (2015), A859 –A889 (cit. on p. 111).

[215] Y. Saad. "A flexible inner-outer preconditioned GMRES algorithm". In: *SIAM J. Sci. Statist. Comput.* 14.2 (1993), pp. 461–469 (cit. on pp. 60, 63).

[216] Y. Saad. "Analysis of augmented Krylov subspace methods". In: *SIAM J. Matrix Anal. Appl.* 18 (1997), pp. 435–449 (cit. on p. 63).

[217] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Second edition. SIAM, Philadelphia, 2003 (cit. on pp. 1, 3, 59, 60, 76, 78).

[218] Y. Saad and M. H. Schultz. "GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems." In: *SIAM J. Sci. Statist. Comput.* 7 (1986), pp. 856–869 (cit. on p. 16).

[219] M. V. Sarkis. "Schwarz Preconditioners for Elliptic Problems with Discontinuous Coefficients Using Conforming and Non-Conforming Elements". PhD thesis. Courant Institute of Mathematical Sciences, Department of Mathematics, 1994 (cit. on pp. 38, 53).

[220] C. Schwab. *p– and hp– Finite Element Methods*. Oxford Science Publications, 1998 (cit. on pp. 33, 34, 36).

[221] C. Schwab and M. Suri. "The p and hp version of the finite element method for problems with boundary layers". In: *Math. Comp.* 65 (1996), pp. 1403–1429 (cit. on p. 36).

[222] A. H. Sheikh, D. Lahaye, and C. Vuik. "On the convergence of shifted Laplace preconditioner combined with multilevel deflation". In: *Numer. Linear Algebra Appl.* 20 (2013), pp. 645–662 (cit. on pp. 10, 15).

[223] V. Simoncini and D. B. Szyld. "Flexible inner-outer Krylov subspace methods". In: *SIAM J. Numer. Anal.* 40 (2003), pp. 2219–2239 (cit. on p. 18).

[224] V. Simoncini and D. B. Szyld. "Recent computational developments in Krylov subspace methods for linear systems". In: *Numer. Linear Algebra Appl.* 14 (2007), pp. 1–59 (cit. on pp. 3, 18, 57, 59).

[225] G. L. G. Sleijpen and H. A. Van der Vorst. "A Jacobi–Davidson iteration method for linear eigenvalue problems". In: *SIAM J. Matrix Anal. Appl.* 17.2 (1996), pp. 401–425 (cit. on p. 63).

[226] B. F. Smith, P. E. Bjørstad, and W. D. Gropp. *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge University Press, 1996 (cit. on pp. 2, 3, 34, 39, 40, 42, 48, 53–55, 57).

[227] P. Solin, K. Segeth, and I. Dolezel. *Higher-Order Finite Element methods*. Chapman & Hall/CRC Press, 2003 (cit. on p. 34).

[228] F. Sourbier, S. Operto, J. Virieux, P. Amestoy, and J. Y. L' Excellent. "FWT2D : a massively parallel program for frequency-domain Full-Waveform Tomography of wide-aperture seismic data - Part 1: algorithm". In: *Computer & Geosciences* 35 (2009), pp. 487–495 (cit. on p. 13).

[229] F. Sourbier, S. Operto, J. Virieux, P. Amestoy, and J. Y. L' Excellent. "FWT2D : a massively parallel program for frequency-domain Full-Waveform Tomography of wide-aperture seismic data - Part 2: numerical examples and scalability analysis". In: *Computer & Geosciences* 35 (2009), pp. 496–514 (cit. on p. 13).

[230] N. Spillane, V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl. "Abstract robust coarse spaces for systems of PDEs via Gereralized Eigenproblems in the Overlap". In: *Numer. Math.* 126.4 (2014), pp. 741–770 (cit. on p. 97).

[231] J. Steiner, D. Ruprecht, R. Speck, and R. Krause. "Convergence of Parareal for the Navier-Stokes equations depending on the Reynolds number". In: *Numerical Mathematics and Advanced Applications - ENUMATH 2013*. Ed. by Assyr Abdulle, Simone Deparis, Daniel Kressner, Fabio Nobile, and Marco Picasso. Vol. 103. Lecture Notes in Computational Science and Engineering. Springer International Publishing, 2015, pp. 195–202 (cit. on p. 108).

[232] C. Stolk. "A rapidly converging domain decomposition method for the Helmholtz equation". In: *J. Comp. Phys.* 241 (2013), pp. 240–252 (cit. on p. 10).

[233] C. Stolk, M. Ahmed, and S. K. Bhowmik. "A multigrid method for the Helmholtz equation with optimized coarse grid correction". In: *SIAM J. Sci. Comput.* 36 (2014), A2819–A2841 (cit. on pp. 12, 13, 15).

[234] K. Stüben and U. Trottenberg. "Multigrid methods: fundamental algorithms, model problem analysis and applications". In: *Multigrid methods, Koeln-Porz, 1981, Lecture Notes in Mathematics, volume 960*. Ed. by W. Hackbusch and U. Trottenberg. Springer-Verlag, 1982 (cit. on pp. 3, 8, 14, 19, 23).

[235] E. de Sturler. "Nested Krylov methods based on GCR". In: *J. Comput. Appl. Math.* 67.1 (1996), pp. 15–41 (cit. on pp. 71, 72).

[236] E. de Sturler. "Truncation strategies for optimal Krylov subspace methods". In: *SIAM J. Numer. Anal.* 36.3 (1999), pp. 864–889 (cit. on pp. 71, 72).

[237] X. Sun and C. Bischof. "A basis-kernel representation of orthogonal matrices". In: *SIAM J. Matrix Anal. Appl.* 16 (1995), pp. 1184–1196 (cit. on p. 78).

[238] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Philadelphia, 2005 (cit. on p. 11).

[239] C. A. Thole and U. Trottenberg. "Basic smoothing procedures for the multigrid treatment of elliptic 3D operators". In: *Appl. Math. Comput.* 19 (1986), pp. 333–345 (cit. on p. 19).

[240] A. Toselli and X. Vasseur. "A numerical study on Neumann-Neumann and FETI methods for *hp*-approximations on geometrically refined boundary layer meshes in two dimensions". In: *Comput. Methods Appl. Mech. Engrg.* 192 (2003), pp. 4551–4579 (cit. on pp. 43, 47–50).

[241] A. Toselli and X. Vasseur. "A numerical study on Neumann-Neumann methods for *hp* approximations on geometrically refined boundary layer meshes II: Three-dimensional problems". In: *M2AN* 40.1 (2006), pp. 99–122 (cit. on pp. 41, 42, 55).

[242] A. Toselli and X. Vasseur. *Domain decomposition methods of Neumann-Neumann type for hp-approximations on geometrically refined boundary layer meshes in two dimensions*. Tech. rep. 02–15. Seminar für Angewandte Mathematik, ETH Zürich, Switzerland, 2002 (cit. on pp. 36, 39, 46, 54).

[243] A. Toselli and X. Vasseur. "Domain decomposition preconditioners of Neumann-Neumann type for *hp*-approximations on boundary layer meshes in three dimensions". In: *IMA J. Numer. Anal.* 24.1 (2004), pp. 123–156 (cit. on pp. 36, 38, 39, 43, 44, 46, 50, 54).

[244] A. Toselli and X. Vasseur. "Dual-primal FETI algorithms for edge element approximations: two-dimensional h and p finite elements on shape-regular meshes". In: *SIAM J. Numer. Anal.* 42.6 (2005), pp. 2590–2611 (cit. on p. 34).

[245] A. Toselli and X. Vasseur. "Robust and efficient FETI domain decomposition algorithms for edge element approximations". In: *COMPEL* 24.2 (2005), pp. 396–407 (cit. on pp. 34, 55).

[246] A. Toselli and O. Widlund. *Domain Decomposition methods - Algorithms and Theory*. Springer, Berlin, 2005 (cit. on pp. 2, 3, 9, 34, 39, 48, 53–55, 57, 95, 96).

[247] U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid*. Academic Press Inc., London, 2001 (cit. on pp. 3, 8, 21, 23, 24, 26, 89–92).

[248] E. Turkel, D. Gordon, R. Gordon, and S. Tsynkov. "Compact 2D and 3D sixth order schemes for the Helmholtz equation with variable wavenumber". In: *J. Comp. Phys.* 232 (2013), pp. 272–287 (cit. on p. 12).

[249] N. Umetani, S. P. McLachlan, and C. W. Oosterlee. "A multigrid-based shifted Laplacian preconditioner for fourth-order Helmholtz discretization". In: *Numer. Linear Algebra Appl.* 16 (2009), pp. 603–626 (cit. on p. 13).

[250] W. Vanroose, B. Reps, and H. bin Zubair. *A polynomial multigrid smoother for the iterative solution of the heterogeneous Helmholtz problem.* Technical Report. http://arxiv.org/abs/1012.5379. University of Antwerp, Belgium, 2010 (cit. on p. 16).

[251] X. Vasseur. *A FMG-FAS procedure for the fully coupled resolution of the Navier-Stokes equations on cell-centered colocated grids.* Talk given at the 1997 Copper Mountain Conference on Multigrid Methods, Copper Mountain, Colorado, USA. 1997 (cit. on p. 8).

[252] X. Vasseur. "Etude numérique de techniques d'accélération de convergence lors de la résolution des équations de Navier-Stokes en formulation découplée ou fortement couplée". PhD thesis. Université de Nantes, France, 1998 (cit. on pp. 8, 105).

[253] P. S. Vassilevski. *Multilevel Block Factorization Preconditioners, Matrix-based Analysis and Algorithms for Solving Finite Element Equations.* Springer, New York, 2008 (cit. on pp. 1, 18).

[254] A. Vion and C. Geuzaine. "Double sweep preconditioner for optimized Schwarz methods applied to the Helmholtz problem". In: *J. Comp. Phys.* 266 (2014), pp. 171–190 (cit. on p. 10).

[255] J. Virieux, S. Operto, H. Ben Hadj Ali, R. Brossier, V. Etienne, F. Sourbier, L. Giraud, and A. Haidar. "Seismic wave modeling for seismic imaging". In: *The Leading Edge* 25.8 (2009), pp. 538–544 (cit. on p. 15).

[256] B. Vital. "Etude de quelques méthodes de résolution de problème linéaire de grande taille sur multiprocesseur". PhD thesis. Université de Rennes, France, 1990 (cit. on pp. 77, 78).

[257] H. A. van der Vorst. *Iterative Krylov Methods for Large Linear Systems.* Cambridge University Press, Cambridge, 2003 (cit. on pp. 1, 3, 59).

[258] H. A. van der Vorst and C. Vuik. "GMRESR: A family of nested GMRES methods". In: *Numer. Linear Algebra Appl.* 1 (1994), pp. 369–386 (cit. on p. 71).

[259] S. Wang, M. V. de Hoop, and J. Xia. "Acoustic inverse scattering via Helmholtz operator factorization and optimization". In: *J. Comp. Phys.* 229 (2010), pp. 8445–8462 (cit. on pp. 9, 13).

[260] S. Wang, M. V. de Hoop, and J. Xia. "On 3D modeling of seismic wave propagation via a structured parallel multifrontal direct Helmholtz solver". In: *Geophysical Prospecting* 59 (2011), pp. 857–873 (cit. on pp. 9, 13).

[261] T. Washio and C. W. Oosterlee. "Krylov subspace acceleration for nonlinear multigrid schemes". In: *Electron. Trans. Numer. Anal.* 6 (1997), pp. 271–290 (cit. on p. 105).

[262] A. Wathen. "Preconditioning". In: *Acta Numer.* 24 (2015), pp. 329–376 (cit. on p. 2).

[263] C. Weisbecker. "Improving Multifrontal Solvers by Means of Algebraic Block Low-Rank Representations". PhD thesis. Institut National Polytechnique de Toulouse, France, 2013 (cit. on p. 9).

[264] R. Wienands and C. W. Oosterlee. "On three-grid Fourier analysis for multigrid". In: *SIAM J. Sci. Comput.* 223.2 (2001), pp. 651–671 (cit. on p. 91).

[265] R. Wienands, C. W. Oosterlee, and T. Washio. "Fourier analysis of GMRES(m) preconditioned by multigrid". In: *SIAM J. Sci. Comput.* 22 (2000), pp. 582–603 (cit. on p. 23).

[266] S. Williams, A. Waterman, and D. Patterson. "Roofline: an insightful visual performance model for multicore architectures". In: *Communications ACM* 55.6 (2012), pp. 121–130 (cit. on p. 110).

[267] J. Xu. "Iterative methods by space decomposition and subspace correction". In: *SIAM J. Sci. Comput.* 34 (1992), pp. 581–613 (cit. on p. 3).

[268] J. Xu and L. Zikatanov. "On an energy minimizing basis for algebraic multigrid methods". In: *Comput. Visual. Sci.* 7 (2004), pp. 121–127 (cit. on p. 91).

[269] P. M. De Zeeuw. "Matrix-dependent prolongations and restrictions in a blackbox multigrid solver". In: *J. Comput. Appl. Math.* 33 (1990), pp. 1–27 (cit. on p. 15).

[270] L. Zepeda-Núñez and L. Demanet. "The method of polarized traces for the 2D Helmholtz equation". In: *J. Comp. Phys.* 308 (2016), pp. 347 –388 (cit. on pp. 11, 31).

Résumé

Résumé Les méthodes multigrille et de décomposition de domaine constituent des méthodes efficaces pour la résolution numérique des problèmes issus de la discrétisation de certaines équations aux dérivées partielles intervenant dans de multiples applications en sciences de l'ingénieur. Ce manuscrit couvre quelques aspects récents à propos de ces méthodes itératives destinées à la résolution de tels problèmes conduisant généralement à des systèmes linéaires ou non-linéaires de très grande taille. Plus spécifiquement, nous abordons le cas des méthodes multigrille géométriques, des méthodes de décomposition de domaine sans recouvrement et des méthodes de Krylov en insistant sur leur combinaison. Dans une première partie, la combinaison de méthodes multigrille et de méthodes de Krylov est ainsi illustrée autour de la résolution d'une équation aux dérivées partielles dite d'Helmholtz modélisant les phénomènes de propagation d'ondes dans un milieu hétérogène. Dans une deuxième partie, nous nous concentrons sur une classe de méthodes de décomposition de domaine dans le cadre d'une discrétisation éléments finis de type hp, où le raffinement est autorisé en diminuant le pas de maillage h ou en augmentant le degré polynômial d'approximation p sur chaque élément. Des résultats théoriques décrivant le comportement des nombres de conditionnement de l'opérateur préconditionné sont donnés et illustrés sur des problèmes académiques. Dans une troisième partie, nous passons en revue des avancées récentes concernant les méthodes de Krylov autorisant l'emploi de préconditionnements variables. Nous détaillons notamment les méthodes de Krylov flexibles munies d'augmentation ou de déflation, où la déflation vise à capturer de l'information de type sous-espace invariant approché. Ensuite, nous présentons des méthodes de Krylov flexibles pour la résolution de systèmes à multiples seconds membres donnés simultanément. L'efficacité des méthodes proposées est illustrée sur des applications frontières en géophysique, nécessitant la résolution de systèmes linéaires de très grande taille sur calculateurs massivement parallèles. Enfin, ce manuscrit se conclut par une évocation des pistes de recherche du candidat dans un futur proche à propos de l'analyse et du développement de méthodes efficaces pour la résolution numérique des équations aux dérivées partielles sur machines massivement parallèles.

Mots-clés Balancing Neumann-Neumann (BNN); Calcul à haute performance; Déflation spectrale; Equations aux dérivées partielles; Equation d'Helmholtz; Equations de Navier-Stokes; Full Approximation Scheme (FAS); Finite Element Tearing and Interconnecting (FETI); Full Multigrid (FMG); Méthode de décomposition de domaine sans recouvrement; Méthode de Krylov; Méthode de Krylov par bloc; Méthode de sous-structuration; Méthode éléments finis de type hp; Méthode itératives; Méthode multigrille; Préconditionnement; Préconditionnement variable; Systèmes linéaires avec multiples seconds membres.